Pervasive Technology Institute Annual Report: Research Innovations and Advanced Cyberinfrastructure Services in Support of IU Strategic Goals during FY 2017

Craig A. Stewart
Beth Plale
Von Welch
Marlon Pierce
Geoffrey Fox
Thomas G. Doak
David Y. Hancock
Robert Henschel
Matthew R. Link
Therese Miller
Eric A. Wernert
Michael J. Boyles
Ben Fulton
Le Mai Weakley
Robert J. Ping
Tassie Gniady
Winona Snapp-Childs

Indiana University
PTI Technical Report PTI-TR17-010
Last revised July 31, 2017

Citation:
Stewart, C.A., Plale, B., Welch, V., Pierce, M., Fox, G., Hancock, D.Y., Henschel, R., Link, M.R., Miller, T., Wernert, E.A., Boyles M.J., Fulton, B., Weakley, L. M., Ping, R.J., Gniady, T., Snapp-Childs, W. (2017). Pervasive Technology Institute Annual Report: Research Innovations and Advanced Cyberinfrastructure Services in Support of IU Strategic Goals During FY 2017. PTI Technical Report PTI-TR17-010. Retrieved from http://hdl.handle.net/2022/21809

Executive Summary

The Pervasive Technology Institute (PTI) is enabled by collaborations across Office of the Vice President for Information Technology (OVPIT), University Information Technology Services (UITS), the IU Maurer School of Law, the IU School of Informatics and Computing, the College of Arts and Sciences, and the Kelley School of Business. The PTI centers are:

- Center for Applied Cybersecurity Research (CACR), led by Von Welch
- Science Gateways Research Center, led by Dr. Marlon Pierce
- Data to Insight Center, led by Professor of Informatics Dr. Beth Plale, also PTI Science Director
- Digital Science Center, led by Distinguished Professor of Computer Science Dr. Geoffrey C. Fox
- National Center for Genome Analysis Support, led by Dr. Thomas G. Doak
- Research Technologies Division of UITS, led by Associate Dean Dr. Craig A. Stewart, also Executive Director of PTI. Below are the units within Research Technologies.
    - Advanced Cyberinfrastructure: High Performance Systems (HPS), Jetstream
    - Community Engagement and Interoperability: Campus Bridging and Research Infrastructure (CBRI), Collaboration and Engagement Support (CESG), Jetstream Project Management and Outreach
    - Science Community Tools: Advanced Biomedical IT Core (ABITC), Advanced Parallel Applications, Scalable Compute Archive (SCA), Science Applications and Performance Tuning (SciAPT)
    - Systems: Research Storage, Application Desktop Virtualization, High Performance File Systems (HPFS), High Throughput Computing (HTC)
    - Visualization and Analytics: Advanced Visualization Laboratory (AVL), Research Analytics, Research Data Services, Digital Humanities Cyberinfrastructure (CyberDH)

In FY 2017, the PTI maintains its sustainability, continuing and refining its activities within IU, the state of Indiana, and the nation. Some of the key accomplishments are:

- Department of Homeland Security Software Assurance Marketplace (SWAMP), lead by CACR, launched a new version of its continuous assurance technologies (called “SWAMP-in-a-Box”) that allows for the deployment *on premise* instances of the SWAMP.
- NCGAS aided researchers in completing many *de novo* assemblies and genomic analysis including that of the endangered crawfish frog.
- SGRC receives significant funding from the National Science Foundation for the development of science gateways.
- Jetstream, the 1st production cloud funded by the National Science Foundation (NSF grant #1445604) enters its 1st year of production.
- D2I Director Beth Plale received an award from the National Science Foundation for Robust Persistent Identification of Data (NSF grant #1659310)

In 1999, under the leadership of Michael A. McRobbie (then vice president of information technology), Indiana University (IU) proposed a partnership with Lilly Endowment to fund key elements of the launch of the School of Informatics and Computing and the creation of six advanced information technology labs that became known as the Pervasive Technology Labs (PTL). The core rationale had three main components: Indiana's economy was lagging behind many other states and in fact leading the nation in several negative economic indicators; state leadership had established a strategy of restoring the Indiana economy on the basis of life sciences and information technology; and, with funding to build on and expand existing strengths in computer science and information technology, IU could contribute strongly to economic as well as scientific and societal good in Indiana.

In 2008, IU proposed the transformation of the Pervasive Technology Labs into the Pervasive Technology Institute (PTI) , enabled by a second round of funding from the Lilly Endowment and increased IU support. Since that time, the PTI has continued to grow and has achieved sustainability. Here we outline the major activities of the PTI as a whole and for each individual center. We also demonstrate how the PTI has positive impacts at IU, the state of Indiana, the nation, and internationally.

IU Impact

PTI, particularly the cyberinfrastructure facilities delivered and supported by the Research Technologies Division of UITS, is a highlight for Indiana University in recruiting and retaining top faculty talent. The role of PTI cyberinfrastructure is featured in a faculty recruitment video, “Supporting the building blocks of discovery: Indiana University’s Advanced Cyberinfrastructure[1].” Every year, UITS conducts a survey of all systems and services. Services and systems managed by the PTI are consistently highly rated. Below we highlight how our systems and services are viewed by the broader IU community.

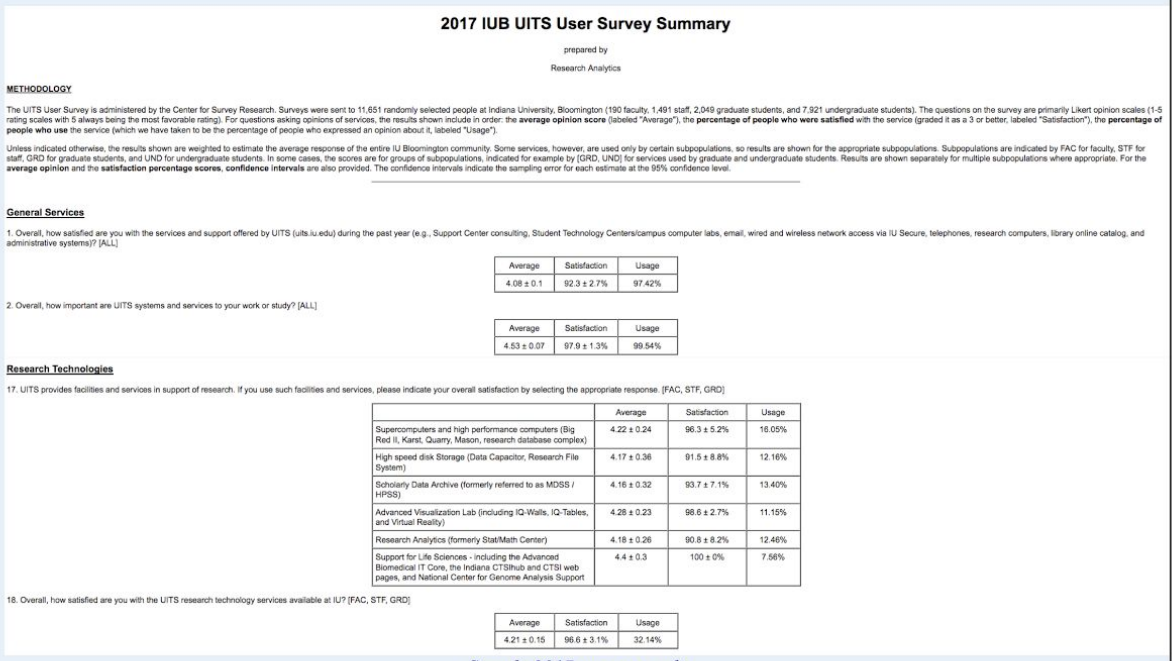
[1] https://www.youtube.com/watch?v=cnlX6uhJVqI

Highlight: PTI services are highly rated by IU faculty, staff, and students

PTI staff are engaged in highly-visible and impactful evaluation and assessment activities at the university and national levels, supporting and informing grant efforts, local service delivery, and national cyberinfrastructure projects. In partnership with the Center for Survey Research, the Office of the Vice Provost for Undergraduate Education on the IUB campus, and officials on IU's regional campuses, University Information Technology Services (UITS), led by staff from its Research Technologies and Client Support divisions, regularly survey the university's faculty, staff, and student stakeholders. Data from this annual survey guides IT leadership in decisions to expand, reduce, and/or make course corrections in existing services; informs and supports decisions to implement new services; and aids decisions makers in anticipating trends and needs on the horizon.

Efforts to increase participation in the survey have resulted in more individuals having a voice in the evolution of IU's IT environment, while also increasing the robustness and veracity of the data used to underpin future planning. We've also made the data collected easier to use by IT leaders and decision makers, employing sentiment analysis methodologies used in social science and humanities research for deeper and more meaningful insights, and creating an interactive dashboard (that allows a more granular and targeted view of the data), making it possible to view data by campus and population.

Again in 2017, Research Technologies' resources and services, which include high performance computing, visualization, analytics, and storage resources, as well as expert consulting and development services, received high marks from its constituents, with its entire suite of services garnering satisfaction scores of 90% or better. Take a peek at the results at: http://www.iu.edu/~uitssur/



2017 IUB UITS User Survey Summary

prepared by

Research Analytics

METHODOLOGY

The UITS User Survey is administered by the Center for Survey Research. Surveys were sent to 11,651 randomly selected people at Indiana University, Bloomington (190 faculty, 1,491 staff, 2,049 graduate students, and 7,921 undergraduate students). The questions on the survey are primarily Likert opinion scales (1-5 rating scales with 5 always being the most favorable rating). For questions asking opinions of services, the results shown include in order: the **average opinion score** (labeled "Average"), the **percentage of people who were satisfied** with the service (graded it as a 3 or better, labeled "Satisfaction"), the **percentage of people who use** the service (which we have taken to be the percentage of people who expressed an opinion about it, labeled "Usage").

Unless indicated otherwise, the results shown are weighted to estimate the average response of the entire IU Bloomington community. Some services, however, are used only by certain subpopulations, so results are shown for the appropriate subpopulations. Subpopulations are indicated by FAC for faculty, STF for staff, GRD for graduate students, and UND for undergraduate students. In some cases, the scores are for groups of subpopulations, indicated for example by [GRD, UND] for services used by graduate and undergraduate students. Results are shown separately for multiple subpopulations where appropriate. For the **average opinion** and the **satisfaction percentage scores**, **confidence intervals** are also provided. The confidence intervals indicate the sampling error for each estimate at the 95% confidence level.

General Services

1. Overall, how satisfied are you with the services and support offered by UITS (uits.iu.edu) during the past year (e.g., Support Center consulting, Student Technology Centers/campus computer labs, email, wired and wireless network access via IU Secure, telephones, research computers, library online catalog, and administrative systems)? [ALL]

Average	Satisfaction	Usage
4.08 ± 0.1	92.3 ± 2.7%	97.42%

2. Overall, how important are UITS systems and services to your work or study? [ALL]

Average	Satisfaction	Usage
4.53 ± 0.07	97.9 ± 1.3%	99.54%

Research Technologies

17. UITS provides facilities and services in support of research. If you use such facilities and services, please indicate your overall satisfaction by selecting the appropriate response. [FAC, STF, GRD]

	Average	Satisfaction	Usage
Supercomputers and high performance computers (Big Red II, Karst, Quarry, Mason, research database complex)	4.22 ± 0.24	96.3 ± 5.2%	16.05%
High speed disk Storage (Data Capacitor, Research File System)	4.17 ± 0.36	91.5 ± 8.8%	12.16%
Scholarly Data Archive (formerly referred to as MDSS / HPSS)	4.16 ± 0.32	93.7 ± 7.1%	13.40%
Advanced Visualization Lab (including IQ-Walls, IQ-Tables, and Virtual Reality)	4.28 ± 0.23	98.6 ± 2.7%	11.15%
Research Analytics (formerly Stat/Math Center)	4.18 ± 0.26	90.8 ± 8.2%	12.46%
Support for Life Sciences - including the Advanced Biomedical IT Core, the Indiana CTSIhub and CTSI web pages, and National Center for Genome Analysis Support	4.4 ± 0.3	100 ± 0%	7.56%

18. Overall, how satisfied are you with the UITS research technology services available at IU? [FAC, STF, GRD]

Average	Satisfaction	Usage
4.21 ± 0.15	96.6 ± 3.1%	32.14%

Sample 2017 survey results

PTI Goal: Create capabilities with which researchers at IU (and beyond) associate and collaborate

PTI supports a number of online information systems used at IU, throughout the academic community as a whole, and by the citizenry of the state of Indiana.

- IUScholarWorks. IUScholarWorks (scholarworks.iu.edu), operated by the IU Libraries, serves as IU's primary persistent digital repository. It is based on a front end that runs under the open source DSpace software. Its back end is the Scholarly Data Archive. IUScholarWorks is the tool IU is using to ensure that the wealth of data and information collected and generated by the IU community remains accessible to and useful for generations to come.
- Indiana CTSI HUB. This is the online portal for the Indiana Clinical and Translational Science Institute, and one of the most widely used data access resources within the IU clinical and translational research community.
- Indiana Spatial Data Portal for GIS data. The ISDP provides access to more than 30 terabytes of Indiana geospatial data. The ISDP provides a variety of geospatial data sets for the state, including the most recent orthophotography and lidar data commissioned by the state. Every year, the ISDP website has thousands of visitors and enables tens of thousands of downloads (through the ISDP multi-file download interface) with total volume of data downloaded exceeding 1 TB.
- ODI-PPA. The ODI Pipeline, Portal, and Archive (ODI-PPA) is a comprehensive web-based solution that provides astronomers and WIYN Consortium members (University of Wisconsin, Indiana University, National Optical Astronomy Observatory, and the University of Missouri) with access to the One Degree Imager (ODI). The modern user interface acts as a single data access point coupled with rich computational and visualization capabilities. It supports scientists in handling complex data sets, while enhancing WIYN's scientific productivity. Most of ODI-PPA is powered by software written by (or integrated by) RT staff, and running on RT hardware including SDA. ODI-PPA has also enabled offshoot Scalable Compute Archive (SCA) projects powered by the Trident microservice and software suite, including EMC-SCA and GCS-SCA.
- Cyberinfrastructure Gateway. The Indiana University Cyberinfrastructure Gateway (CI Gateway) is an online portal designed to centralize information about and access to IU's advanced scholarly and artistic CI. The gateway allows users to find information on current queues, get help, see outages, find information on available software, and transfer and manage data.
- Galaxy portal at IU. The National Center for Genome Analysis Support provides three web-based portals that feature easy-to-use interfaces for genomics researchers to create and execute their own workflows on NCGAS systems. Using the Galaxy web portal environment, NCGAS has created Galaxy portals for IU investigators, NSF-funded life science researchers across the nation, and the Penguin On Demand system for federally-funded investigators. These provide access to the full suite of genome assembly, annotation, alignment, and other applications – as well as file transfer and transformation utilities for building genome science workflows.

One of the major ways in which students and faculty interact with their collaborators and exchange ideas is through conferences and workshops. PTI proudly supports such endeavors and frequently hosts such events at IU.

Academic conferences and workshops held at IU that bring scholars here from other institutions						
Conference	Topic	IU attendees (faculty)	IU attendees (non-faculty)	IU attendees (total)	Non-IU faculty attendees	Total attendees
LUG 2017	The Lustre User Group is a meeting currently sponsored by the OpenSFS non-profit which brings together users and developers of the open source Lustre file system. IU staff led the overall event organization chairing both the program and organizing committees.	3	17	20	149	
Sustainability of Coffee Production in Columbia	NCGAS organized a meeting between major coffee research group Cenicafe and collaborators at Cornell	1	3	4	6	
Research Services Expo and Peebles Lecture	Research Expo introduced IU faculty, staff and students to the variety of research services available to the IU community via demos and consulting by RT personnel. The Peebles lecture presented by Felix Bachmann focused on software engineering.					
TOTAL		4	20	24	155	0

Highlights of collaborations at IU and conferences held at IU include the following:

Digital History and the Old Oaken Bucket

Since 1925, the Old Oaken Bucket has been the symbol of a college football rivalry between the Hoosiers of Indiana University and the Boilermakers of Purdue University as it is passed back and forth between the two schools. Coming from a small farm in Southern Indiana, the bucket contains a 40-lb chain made up of letters "I", "P", and "IP". A new bronze letter is added to the chain each year with the score, date and the city where the game was played engraved on the link.

During the Spring 2017 semester, the Bucket became the focus for IU student Gregory Simon's Digital History course project. In collaboration with the Cyberinfrastructure for the Digital Humanities team and the Advanced Visualization Lab, Simon's project features a 3D representation of the Old Oaken Bucket created through photogrammetry and structured light scanning digitization methods (along with some meticulous model editing thanks to AVL intern, Tyler Jackson).

3D print of the Old Oaken Bucket edited, printed, and assembled by Jeff Rogers and Tyler Jackson of the Advanced Visualization Lab

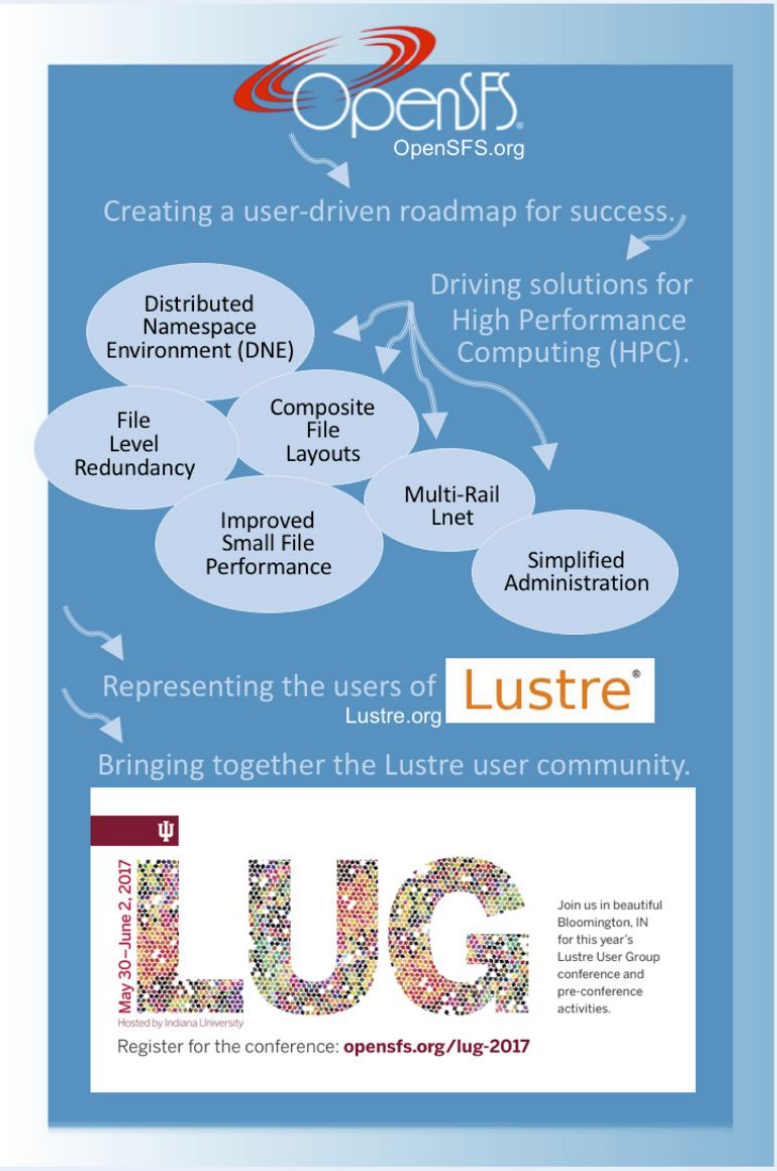
The Digital History course, taught by Kalani Craig, Department of History, explores the rich history of Indiana University—Bloomington with digital tools like text mining, network analysis, and makerspace technology. "The title of the course really says it all." Dr. Craig explains. "The history part is about getting hands-on in a discipline that lets you dig deep into the past of a community you really care about, to see how the local people and movements that shaped IU's present are part of a bigger global past. That's where history's compatibility with digital tools--text mining, network analysis, and so on--can really make a difference for an undergraduate research project. Part of it is that we're learning practical skills and talking to audiences outside the classroom, but the more important thing is that we're looking critically at how these computational tools fit into our work as historians. We're using, but we're also being critical of, the digital tools that increasingly mediate our daily lives, and that means working with IU's past helps us figure out where we fit into IU's digital present." View the 3D model of Old Oaken Bucket and Simon's accompanying explanation of the Bucket's significance here.

Highlight: LUG 2017 held at IU

The Lustre® User Group (LUG) conference is the industry's primary venue for discussion and seminars on the Lustre parallel file system and other open source file system technologies. LUG 2017 was held in Bloomington, Indiana May 30, 2017 – June 2, 2017.

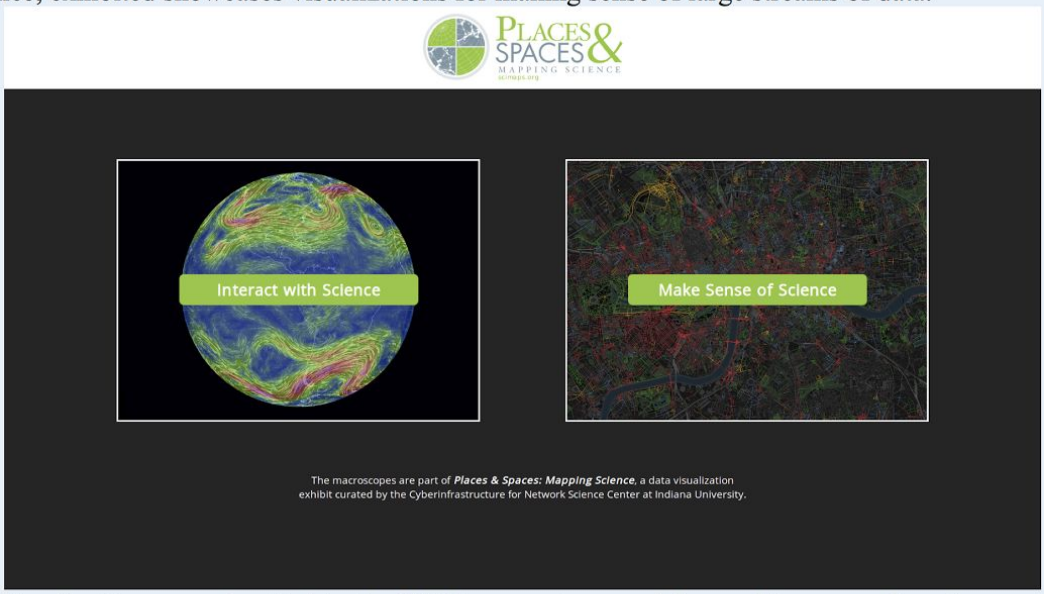
This conference provided attendees with the opportunity to:

- Hear from the world's leading developers, administrators, solution providers, and users of Lustre
- Be an active participant in industry dialogue on best practices and emerging technologies
- Explore upcoming developments of the Lustre file system
- Immerse in the strong Lustre community, working collaboratively to further the development of Lustre

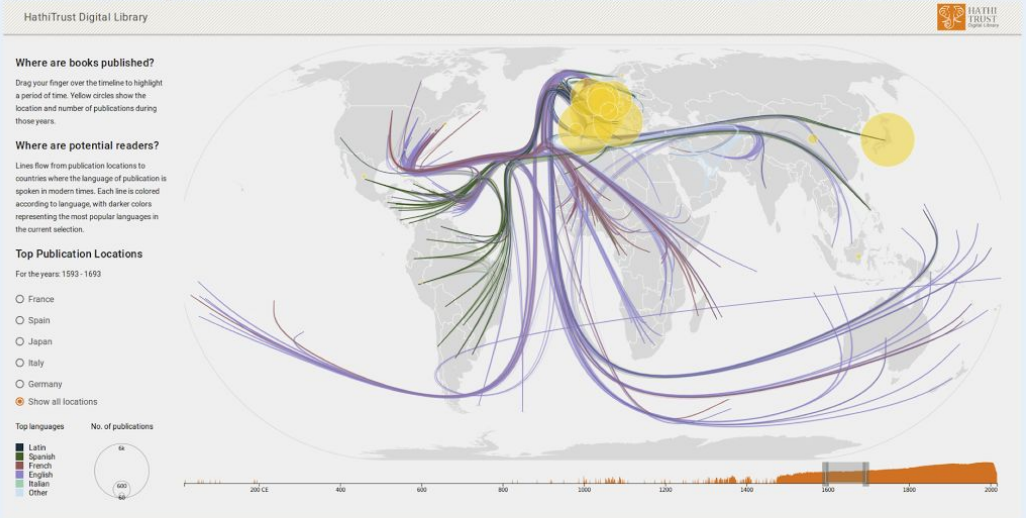


Highlight: Visualizing the HathiTrust texts

The Advanced Visualization Lab (AVL) has an ongoing collaboration with the Cyberinfrastructure for Network Science Center (CNS) to present interactive data visualizations in a multi-touch kiosk. Interactive data visualizations, or macroscopes, have great potential as tools for exploring, understanding, and communicating science. The traveling exhibit, Places & Spaces: Mapping Science, exhibited showcases visualizations for making sense of large streams of data.



The visualizations are submitted to the CNS by experts around the world as examples of "macroscopes" that help the user focus on patterns in data that are too large or complex to see unaided. This year, one of the main visulaizations to make its debut was a library collection mapped in time and space. The library collection was curated by the HathiTrust, led by IU's own Beth Plale. This exhibit was displayed at Vanderbilt University January 23, 2017 – April 23, 2017..



HathiTrust visualization example.

Contributions of PTI overall to IU sponsored research

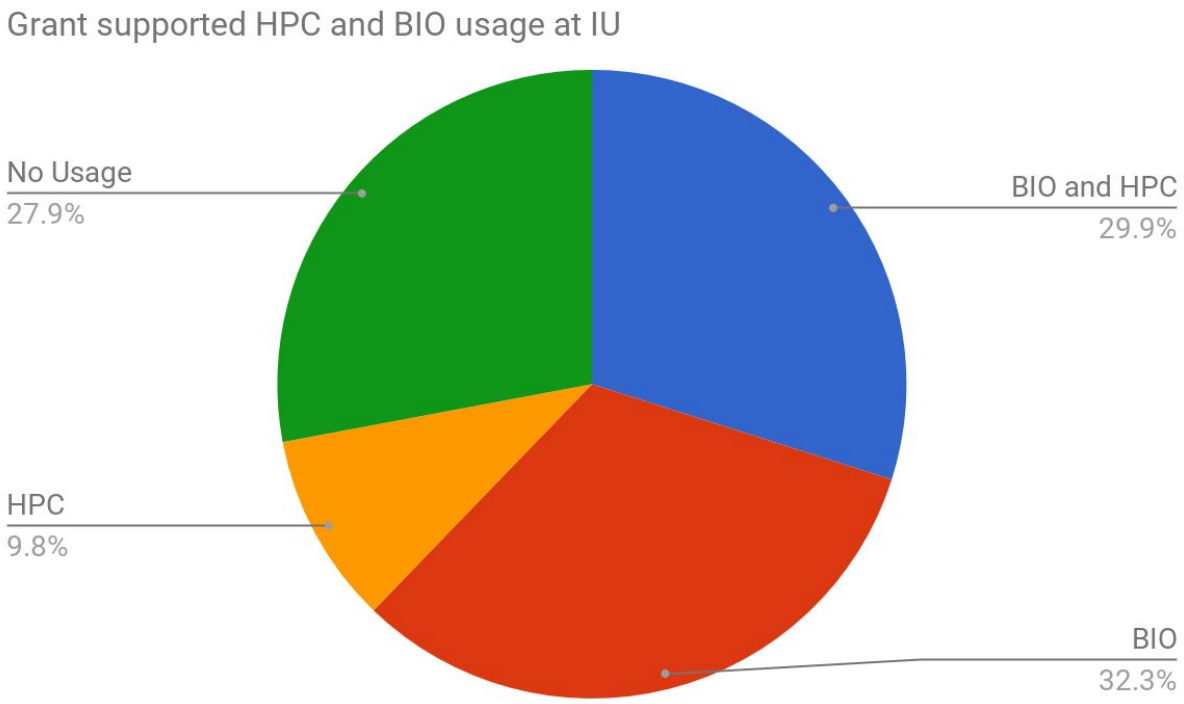
The table below shows awards to each PTI Center, as well as grant awards to IU where a PTI Center is a formal partner. In addition to this sort of partnering, the Research Technologies Division of UITS offers a wide variety of services that are used by researchers in many disciplines and many responsibility centers at IU.

Amount of sponsored research for PTI from inception of PTL in 1999	
PTL & IUPUI 1999 - 2014	$81,489,841
PTI FY 2015	$10,053,716
PTI FY 2016	$16,546,321
PTI FY 2017	$6,593,486
Total	$114,683,364

Use of PTI / Research Technologies Cyberinfrastructure Systems and Services and External Funding to IU

In FY 2017 IU received a total of $500,037,586 in new extramural grants and contracts. IU's cyberinfrastructure contributed significantly to IU's grant competitiveness. The figure below depicts grants and awards to the entire university, subdividing the awards to IU according to use of IU cyberinfrastructure services by PI or Co-PI. New grant and contract receipts are subdivided into four categories:

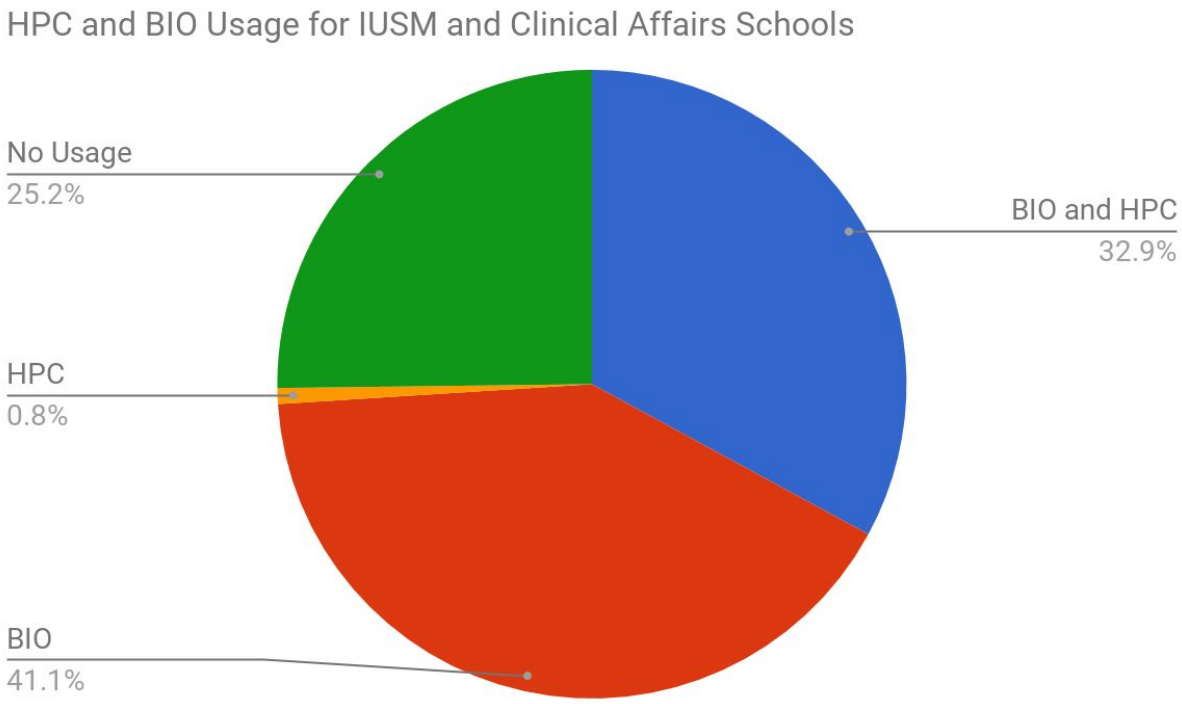
- PI and Co-PIs include no one with an account registered on any CI system or service supported by Research Technologies Division of UITS
- PI and Co-PIs include at least one person with an account on IU's high performance computing (HPC) and cyberinfrastructure services (Big Red II, Quarry, Karst, Mason, Scientific Data Archive, Research File system, etc.)
- PI and Co-PIs include at least one person who uses databases, data resources, or collaboration tools delivered and supported by Research Technologies Division of UITS
- PI and Co-PIs include users of HPC and data /collaboration tools



All in all, more than 70% of grants and awards received by IU in FY 2017 went to a PI / Co-PI team that includes at least one user of cyberinfrastructure resources provided and supported by PTI (particularly the Research Technologies Division of UITS). It is certainly not the case that each and every grant award to a faculty member using PTI resources depended critically on those resources. On the other hand, it seems impossible to believe that these resources are unimportant when a significant sum of research dollars went to people who use these resources. In the past few years, the PTI has taken steps to actively generate return on investment (ROI) measures for IU's cyberinfrastructure[2].

[2]Thota et al. (2016). A PetaFLOPS Supercomputer as a Campus Resource: Innovation, Impact, and Models for Locally-Owned High Performance Computing at Research Colleges and Universities. In Proceedings of the 2016 ACM on SIGUCCS Annual Conference (SIGUCCS '16). ACM, New York, NY, USA, 61-68. DOI: https://doi.org/10.1145/2974927.2974956

The figure below depicts the same sort of grant and award information for the IU School of Medicine and other Clinical Affairs Schools (IU Schools of Dentistry, Health and Rehabilitation Sciences, Nursing, Optometry, Public Health – Bloomington, Richard M. Fairbanks School of Public Health – Indianapolis, Social Work[3]).



PTI Sub-goal: Be the 'partner of choice' within IU and the nation for creating and implementing cyberinfrastructure facilities (particularly when funded by a grant or contract focused on construction of a new facility or delivery of a new capability)

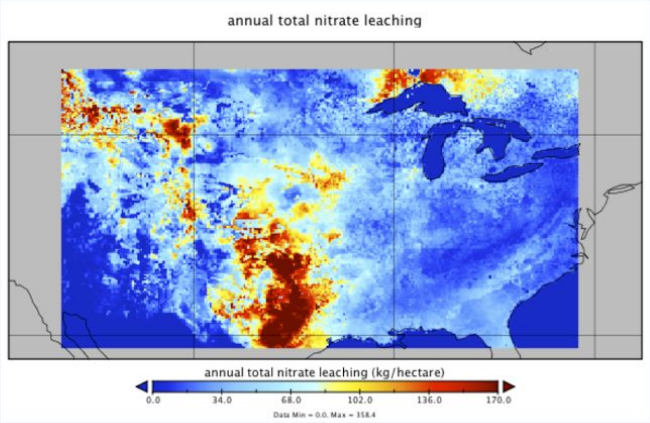
Highlights of PTI activities in creating and implementing cyberinfrastructure facilities includes the following[4]:

[3] http://www.iu.edu/initiatives/clinical-affairs.shtml

[4] Congratulations are due to Dr. Adam Ward who received a National Science Foundation CAREER award (NSF grant #1652293).

Highlight: Improving water quality through advanced simulations

Since the 1950's, there have been considerable shifts in human consumption of goods, climate, and in resource/land management – more intensive farming, more reliance on fertilizers and pesticides. In sum, the United States has shifted from the smallholder family farm in favor of the industrial agriculture style of farming. These changes pose considerable challenges for maintaining high water quality. Adam Ward, Assistant Professor in the IU School of Public and Environmental Affairs, has been conducting exciting new research focused on understanding water quality and water quantity impacts of agricultural activity in the Mississippi River basin, draining more than 1.2 Million square miles of the continental US and Canada.



Nitrate fertilizer losses to groundwater for corn production in 2007.

Experimental DCRAM hardware, provided by Intel.

To do this work, Ward has been collaborating with IU's Science Applications and Performance Tuning (SciAPT) and High Performance File Systems (HPFS) groups to optimize the Agro-IBIS code to run in parallel on IU's Big Red 2 supercomputer using an experimental extreme performance filesystem called DCRAM. This high-throughput, low-latency Lustre file system offers 35 TB of file storage, using a combination of enterprise solid state hard drive (SSD) technology, Infiniband interconnect, and eight state-of-the-art Linux servers. DCRAM, as the smaller but much faster cousin to the DC2 data capacitor, has proven capable of managing the extraordinary I/O requirements of Ward's data-intensive simulations.

This work is helping to provide both a better understanding of the aggregation of activities on water quality and to help guide public policy. Previously, a single model simulation took Ward a week or more to complete on a single processor. Now, with the experimental hardware and optimized code, more than one hundred simulations can be completed in the same amount of time. As a result, new discoveries are coming faster than ever! Ward is now able to complete ensembles of simulations to assess changes in management practices and how they impact water quality and quantity in the Midwestern U.S. This work was supported, in part, by a grant from the National Science Foundation (NSF) Grant #s EAR 1331906 and EAR 1505309, and by Intel – who provided much of the hardware for DCRAM.

Highlight: Patent in Progress for PIPES component

The use of immersive virtual worlds is increasingly widespread in research, development, and applications. Immersive experiences created using virtual worlds are useful in many capacities including education. A virtual reality experience – "PIPES" – developed by Chauncey Frend from the IU Advanced Visualization Lab (AVL). "PIPES" enables virtual reality developers to simulate environmental conditions such as wind, heat, or smell. The experience simulates a guided tour through an ancient Roman dining palace where users can feel warmth from the sun, breezes in the courtyard, and the scents of traditional foods. A video overview of this project can be viewed here PIPES Demo.

The Programmable Immersive Peripheral Environmental System (PIPES) is one of Frend's projects he gets to showcase at major tech conferences around the country

PIPES was first displayed at the Supercomputing conference in 2015, was awarded the "Best Research Demo Award" in 2016 by the IEEE Computer Society and a component of the virtual reality system is currently patent-pending. AVL extended its current virtual reality systems with the PIPES technology to enhance virtual world projects. Existing and future IU personnel benefit from having this new technology and expertise in-house and readily available. "The idea behind PIPES was that we could provide folks who don't do a lot of computer programming or electrical engineering with hardware that they can control with timers or sensors to enhance the user experience," which is how wind, heat, and smell could be incorporated in the ancient Rome reconstruction.

The PIPES control component developed by Frend. The IURTC has managed the patent process and a non-provisional patent has been submitted. #PCT/US2017/018095

PTI Sub-goal: Enable the translation of software innovations to practical use

One of the most notable ways in which the PTI enables the translations of software innovations to practical use is through science gateways. A science gateway is a community-developed set of tools, applications, and data that are integrated via a web portal. Science gateways enable scientists to access high performance computing (HPC) resources without having to be HPC experts. Science gateways are designed specifically to support a particular type of scientific research, with an emphasis on supporting the entire scientific process from start to finish. Science gateways help entire communities of researchers use high-performance computing resources and advanced cyberinfrastructure to pursue common scientific goals.

Highlights of science gateways supported by PTI include the following:

Science Gateways Supported by Science Gateway Research Center			
Name	**Topic**	**Jobs run FY 2017**	**Total Users**
SEAGrid (GRidchem)	The Science and Engineering Applications Grid (SEAGrid.org), formerly known as GridChem, is a science gateway that provides access to computational chemistry, material science, and engineering applications on IU and XSEDE computing infrastructure. SEAGrid is a tenant in the Apache Airavata-based SciGaP hosted services. During FY 2017, SEAGrid was in the top five most heavily used XSEDE gateway by numbers of jobs run and computing hours used.	29,677	925
Geo-Gateway	NASA earthquake researchers to develop science gateways and cyberinfrastructure. This science gateway resulting from this work provides interactive online access to NASA data products and simulation tools. The portal gives users simple access to sophisticated InSAR datasets, GPS time series data, and forward/inversion modeling tools for comparing earthquake fault models. GeoGateway delivers over 3 TB of UAVSAR data (synthetic aperture radar data collected by aircraft) to NASA researchers.	N/A	143
UltraScan	UltraScan science gateway (PI Dr. Borries Demeler, University of Texas Health Science Center San Antonio) allows biophysicists to perform data analysis on analytical ultracentrifugation experiments, uncovering properties of molecules in solution. Through a collaboration with PTI/RT, this data analysis is performed on campus resources, on national cyberinfrastructure (XSEDE), and at international supercomputing centers in Germany. UltraScan is supported by the Apache Airavata-based SciGaP hosted services SGRC.	28,531	~100
Chemcompute	Chemcompute is a science gateway which supports computational chemistry. It enables faculty, graduate students, and undergraduate students to calculate properties of molecules. This science gateway provides a web-based portal that is easy for undergraduate chemistry students to use, eliminating the costly	21,200	1,327

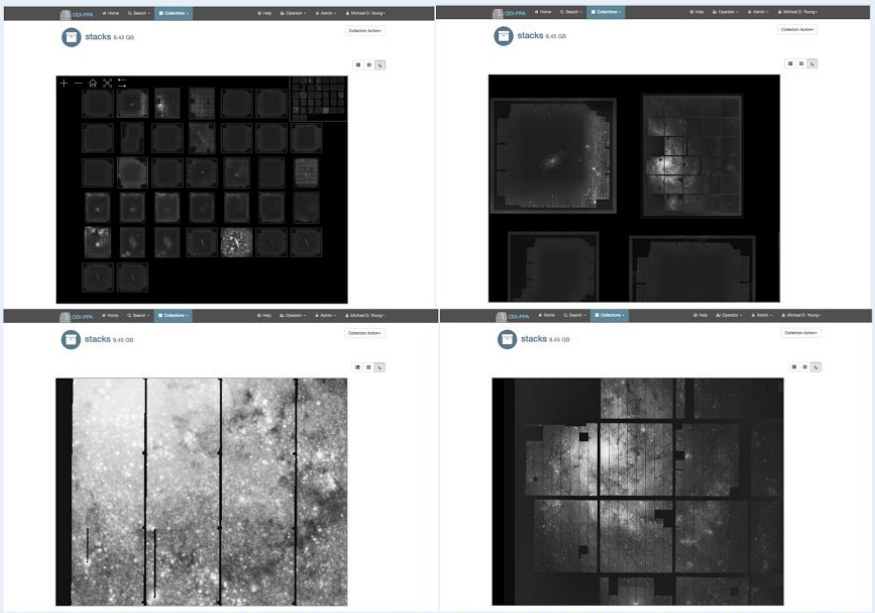
	software.		
University of South Dakota Campus Gateway	This Science Gateway provides access to supercomputing resources without the complexity that often accompanies them. By using this site, users can focus on their computational research instead of worrying about complicated interfaces.	127	11

Other highlights of PTI activities in translation of software innovations to practical use include the following:

- **SPLInter**. The Structural Protein Ligand Interactome is a computational drug design and discovery resource for ranking molecules docked to the human proteome. The portal contains the DOPIN (Docked Proteome Interaction Network) database, which contains millions of pre-docked and pre-scored complexes from thousands of targets from the human proteome and thousands of drug-like small molecules from the NCI diversity set and other sources. SPLInter uses the Open Science Grid for docking simulations and presents visualization, scoring, and ordering information via a web portal.
- **CACR and HTC participate in securing software through the Software Assurance Marketplace (SWAMP).** CACR and the High Throughput Computing Group continue to provide the Software Assurance Marketplace (SWAMP), a $20 million DHS-funded facility that allows software developers and users to more easily identify and fix security vulnerabilities in their software, reducing the risks with using that software.
- **Center for Trustworthy Scientific Cyberinfrastructure.** Led by CACR, the Center for Trustworthy Scientific Cyberinfrastructure consortium is funded by NSF to lead its science community in securing the computational infrastructure critical to today's trustworthy science. In this role, CACR works with software development projects (e.g., Pegasus, SciGaP, Globus, NTP Foundation) to produce more secure software by advising them on good software development practices and the best way to implement new features.

Highlight: Viewing the universe with ImageX

Increasingly, web-based tools and portals are being utilized by scientists and researchers to make new discoveries. The Scalable Compute Archive group in Research Technologies – Pervasive Technology Institute at Indiana University, has been creating such portals for some time. In particular, they developed Image Explorer (ImageX) for astronomical image analysis. In a more recent instantiation, ImageX has been developed for more general-purpose scientific use and can be applied to any scientific image data format convertible to standard formats – thus enabling rapid interactive image visualization for large datasets (of stars down to molecular structures) on a web browser.



Examples shown are from within the ODI-PPA portal for astronomical images taken on the One Degree Imager. Each image is 500 MB - 2 GB, yet ImageX allows rapid interactive visualization of a set of images based on advanced preprocessing.

- **NCGAS continues expanding access to genomic analysis software by researchers.** This includes the NSF-funded core purpose of NCGAS: to give researchers easy access to bioinformatics software packages on HPC systems capable of effectively running them, and to provide these packages in a menu-driven interface with Galaxy. Beyond pure research, NCGAS collaborates on two NIH Information Technologies in Cancer Research grants to make genomics analysis tools available to cancer researchers; the Trinity and GenePattern projects are centered at the Broad Institute and UCSF respectively. NCGAS works to improve these tools, but emphasizes making the tools broadly available.
- **AVL extends and further enhances its Collection Viewer application.** The AVL Collection Viewer software is an application built for collaboratively viewing and interacting with media collection(s). Media can include photos, videos, and audio clips. Commonly used in conjunction with the AVL's IQ-Table or IQ-Tilt systems, the software features an XML configuration file that

allows exhibit creators to quickly and easily tailor the experience. Select new collections for FY17 included a beautiful exhibit featuring IU's Center for Network Science Places and Spaces Mapping Science collection.

- **Open Science Grid Operations hosts network metric datastore and visualization for Large Hadron Collider Computing.** The OSG Operations group implemented a high availability database to collect performance-focused Service Oriented Network monitoring ARchitecture (perfSONAR) metrics for the Worldwide LHC computing facilities. This includes a Cassandra DB system that allows distributed high performance service management. The data collected can be immediately visualized via hosted services to troubleshoot active network issues, or historical data can be used to do long-term analysis of worldwide computing networking.
- **Text analysis with RNotebooks for beginners.** The CyberDH group has adapted algorithms from Matt Jockers into RNotebooks with markup that takes novices through the basics of text analysis (top ten counts, dispersion), working up to corpus clusters (dendograms) and using LDA. The code for this work currently resides on IU GitHub.

PTI Goal: Offer services that enable new innovations and accelerate research by the IU scientific, scholarly, clinical, engineering, and artistic communities

General Use of PTI / Research Technologies Cyberinfrastructure Systems and Services

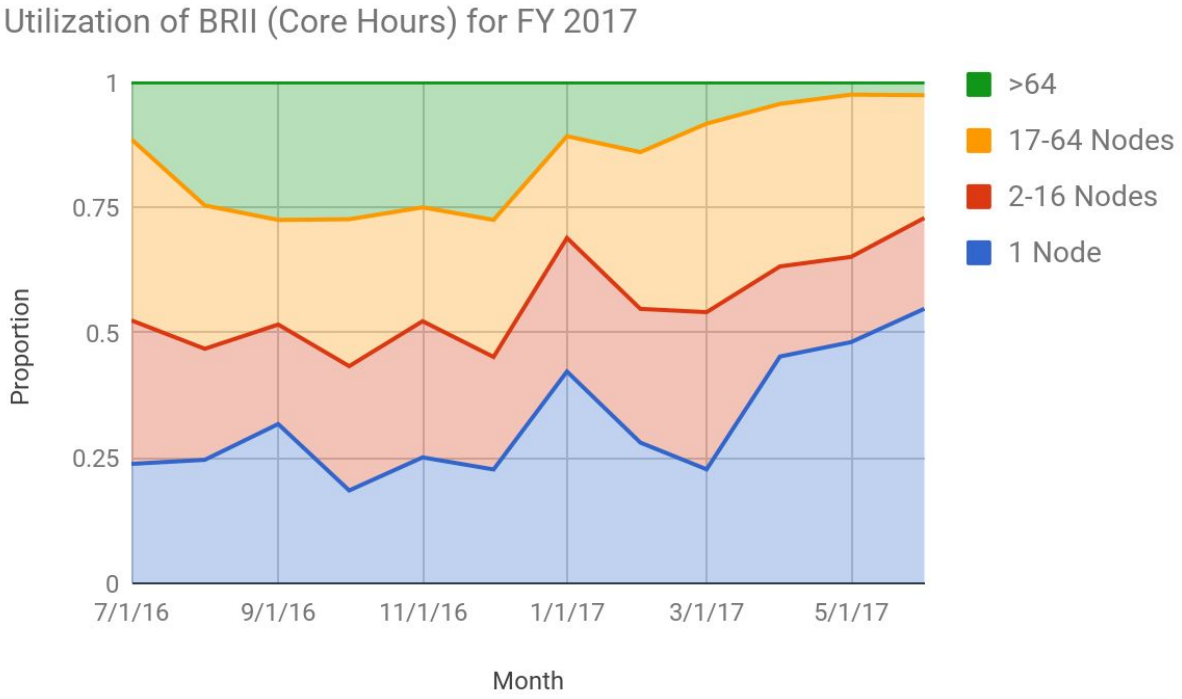
One of the most prominent ways in which the PTI enables new innovations and accelerates research is by providing high performance computing facilities and advanced storage systems. These forms of cyberinfrastructure are both heavily and widely used.

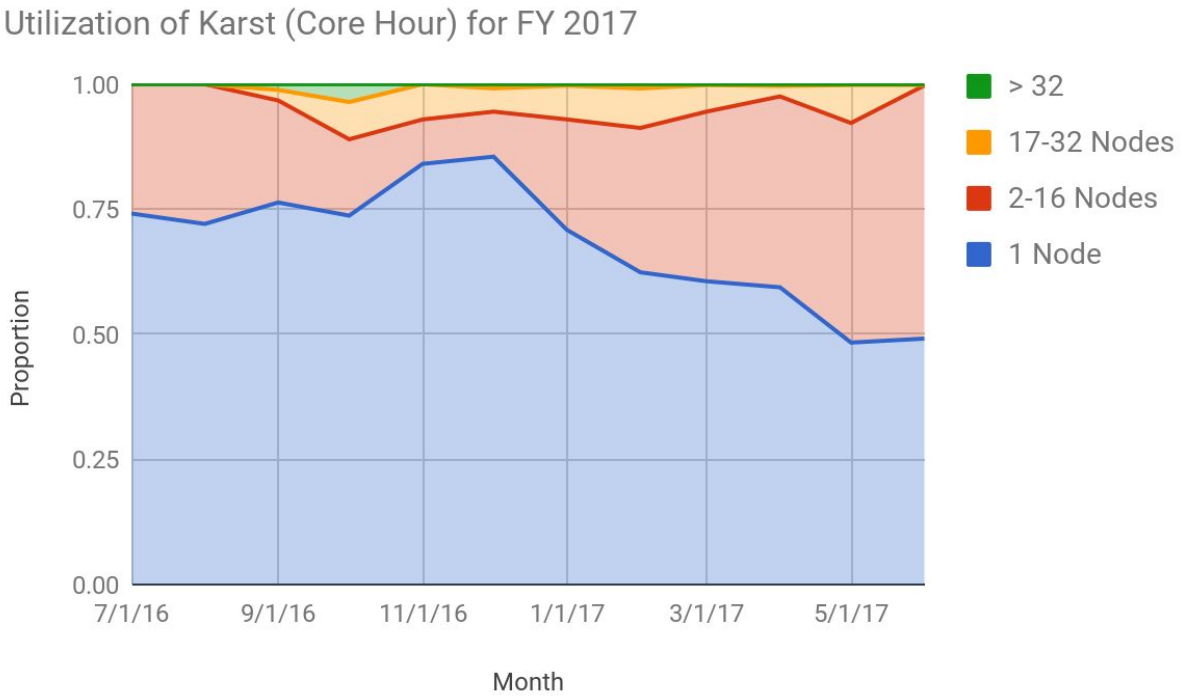
Illustrations of the usage of various systems are portrayed in the figures and tables below.

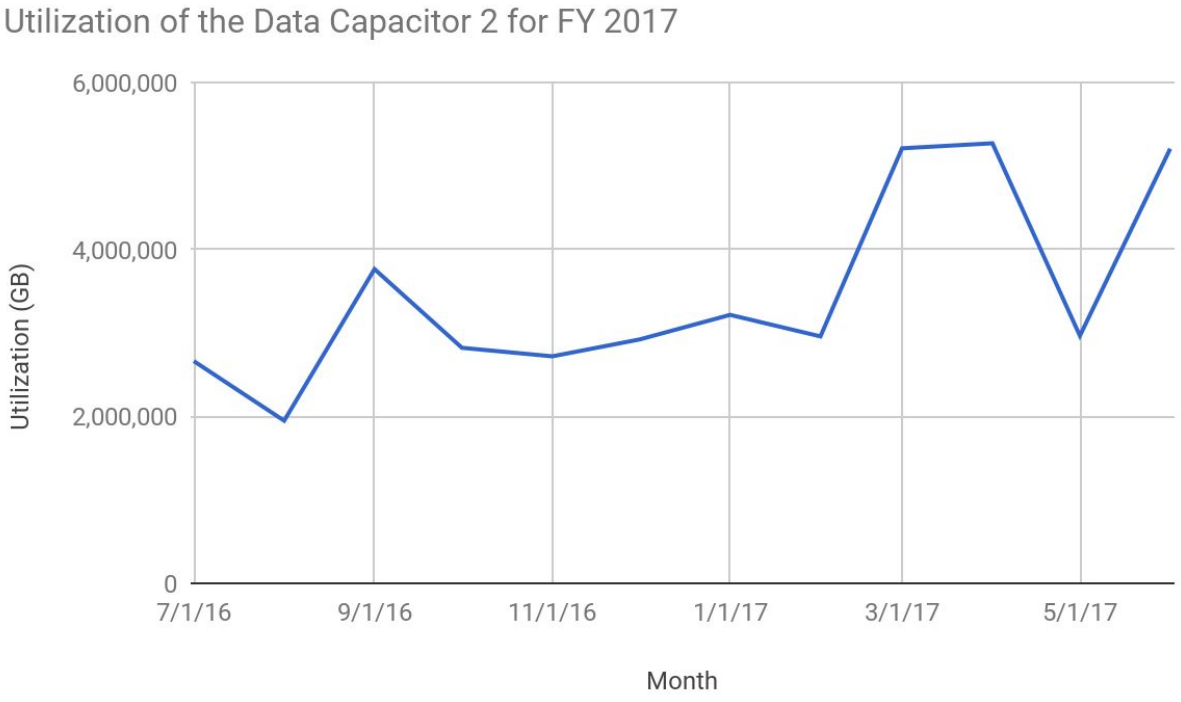
Utilization of Resources						
	Jobs			CPU/Resource hours		
Systems	FY 2015	FY 2016	FY 2017	FY 2015	FY 2016	FY 2017
Big Red II	319,106	397,765	856,567	128,239,686	121,144,818	98,790,001
Karst	424,329	844,529	462,942	17,025,544	21,633,993	23,712,524
Mason	70,648	77,495	68,986	5,961,997	3,168,193	2,969,123
Quarry	3,814,570	Decommissioned	--	5,200,548	Decommissioned	--
XSEDE use by IU researchers	1,277	7,530	31,834	195,664	450,750	6,835,296
Open Science Grid use by IU researchers	25,808,916	27,164,520		26,619,180	27,010, 916	

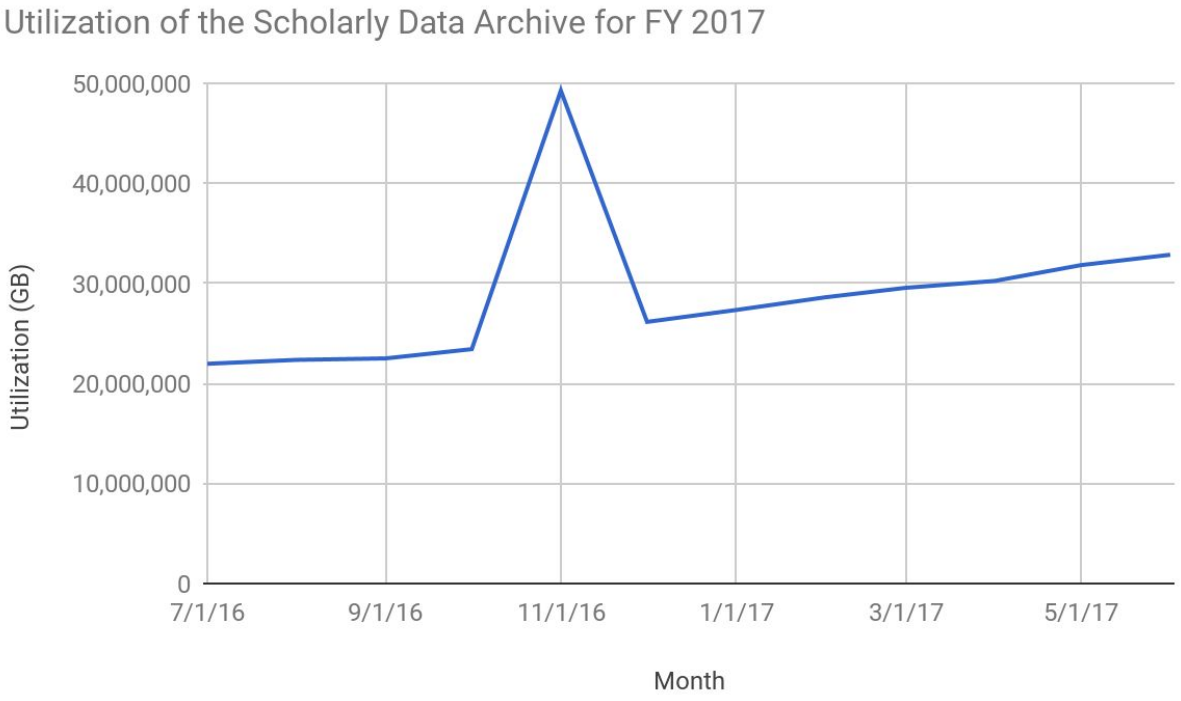
Open Science Grid use by non-IU researchers						1,518,742,000
Total	30,438,846	28,491,839		183,242,619	146,397,756	

	Number of records					Services delivered
Service	**FY 2013**	**FY 2014**	**FY2015**	**FY2016**	**FY 2017**	**Description**
National Gene Vector Biorepository	107,576	111,550	124,588	146,484	238,206	Number of database records
INResearch	1,992	3,367	4,169	5,038	5,986	Completed health profiles in database
REDCap	681	935	1,141	1,562	2,018	New projects using REDCap initiated by researchers
	1,362	2,297	3,438	5,000	7,018	Total projects using REDCap since ABITC assumed responsibility for REDCap in 2010
CTSI Grants Management System	28	25	31	34	36	Grant competitions managed
	106	105	54	66	20	Proposals/submissions awarded
	96	117	146	176	209	Grant competitions managed since inception in 2009
Total Records	155,153	167,997	188,755	258,241	253,493	


Utilization of BRII (Core Hours) for FY 2017
1
0.75
0.5
0.25
0
Proportion
7/1/16
9/1/16
11/1/16
1/1/17
3/1/17
5/1/17
Month
>64
17-64 Nodes
2-16 Nodes
1 Node


Utilization of Karst (Core Hour) for FY 2017
1.00
0.75
0.50
0.25
0.00
Proportion
7/1/16
9/1/16
11/1/16
1/1/17
3/1/17
5/1/17
Month
> 32
17-32 Nodes
2-16 Nodes
1 Node


Utilization of the Data Capacitor 2 for FY 2017
6,000,000
4,000,000
2,000,000
0
Utilization (GB)
7/1/16
9/1/16
11/1/16
1/1/17
3/1/17
5/1/17
Month


Utilization of the Scholarly Data Archive for FY 2017
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
Utilization (GB)
7/1/16
9/1/16
11/1/16
1/1/17
3/1/17
5/1/17
Month

When the UITS Research Technologies Division proposed the purchase of Big Red II, we made a commitment to offer training, information, and support services to increase the diversity of disciplines and sub-disciplines that make use of the system. In particular, we set a goal of having Big Red II used by at least 150 disciplines and sub-disciplines practiced at IU (out of 381 recognized categories). PTI met and exceeded this goal, as shown in the table below.

Disciplines & Sub-disciplines represented among users									
	IUPUI			IUB			Total		
System	FY 2015	FY 2016	FY 2017	FY 2015	FY 2016	FY 2017	FY 2015	FY 2016	FY 2017
Big Red II	68	87	102	141	159	169	159	254	194
Karst	117	152	189	164	218	253	214	180	287
Total	136	168	202	201	238	272	243	279	303

Highlights of of PTI activities in offering services that enable the broader IU scientific, scholarly, clinical, engineering, and artistic communities include the following:

Highlight: '(Re)Imagining Science' exhibition turns collaboration into art

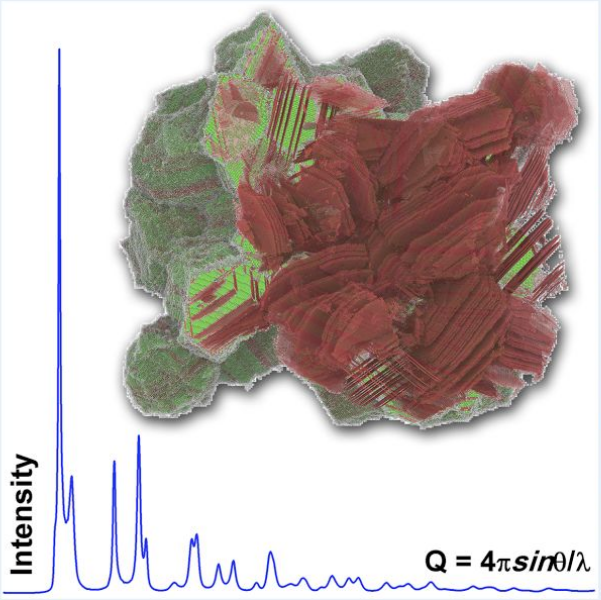
Salamanders in Lily Pad (an example from the collaboration between Margaret Dolinsky and Roger Hangarter)

Artists and research scientists at Indiana University Bloomington teamed up in more than a dozen creative partnerships to visualize scientific principles and foster new ways of understanding. The results of their collaborations were on display in the exhibition "(Re)Imagining Science" October 14 to November 16, 2016 at the Grunwald Gallery. Prominanlty featured was collaboration between Margaret Dolinsky, Advanced Visualization Lab and associate professor of digital art, and Roger Hangarter, Distinguished Professor and Chancellor's Professor of Biology.

In this collaboration, visually stunning line drawing emerge from leaves through the manipulation of chloroplasts. Actual plant leaves are framed for display. The full article appears in IT News & Events and Art at IU.

Highlight: Crystallography methods improved with Big Red II

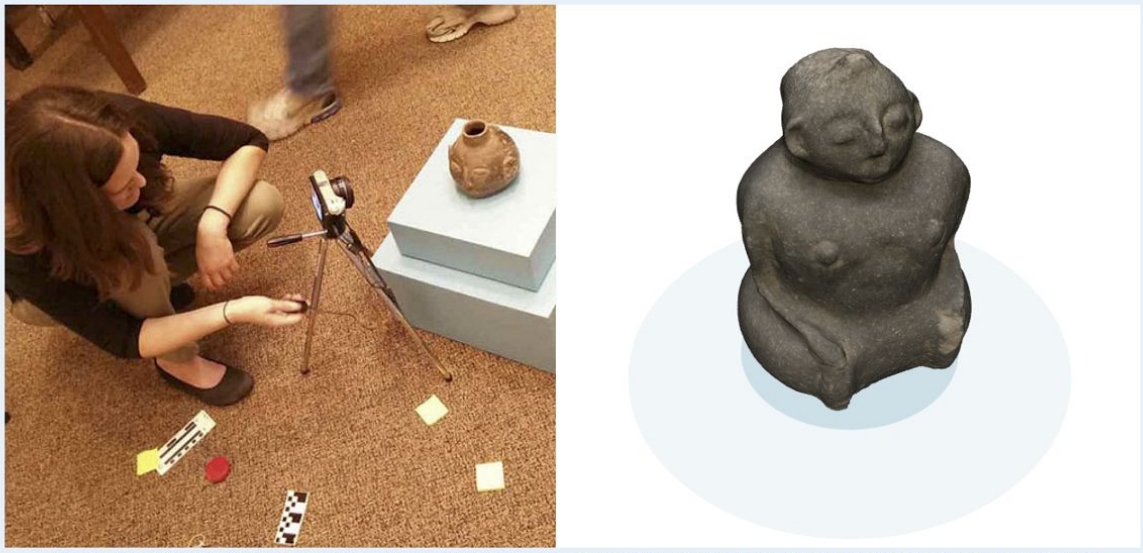
Many materials such as salts, metals, minerals, and biological molecules can form crystals. Understanding the structures and properties of crystals aids in not only a better understanding of nature, it is crucial for the development of new materials and medicines, for example. IU researchers Alberto Leonardi and David Bish, Department of Geological Sciences, have been using Big Red II to do model simulations and develop more sophisticated algorithms to enable the investigation of large collections of microstructures. Leonardi further explained, "Working on the fence, between experimental and theoretical science, my research is oriented to provide more accurate and reliable tools for the development of new applied technologies for applications spanning from business to environmental (i.e., electronics, structural engineering, sustainable energy, environment protection). As the advancement in applied technology and science relay in the accurate understanding of the fundamental behavior of materials so to exploit in the most efficient and environmentally compatible way the availability of these, my research focuses to the study of mechanisms at the atomistic scale that affects the macroscopic behavior of nanostructured materials used for chemical and mechanical applications (i.e., catalysis of chemical reactions in fuel cells, strength and fatigue of engine and vehicle components). In particular, by exploiting the power of computational techniques, my research has allowed an accurate and reliable interpretation of experimental data. The results explored limitations on the current materials characterization techniques, leading to a more comprehensive understanding/knowledge of the role played by the disorder to the directly observable materials behavior."



When asked about the advantages of using high performance computers such as Big Red II to do his research, Leonardi said, "Big Red II provides a very flexible computing environment where to perform highly distributed parallel simulations or solve a large collection of independent computations on CPU, GPU or hybrid CPU+GPU hardware environment. The flexible and roomy memory hardware provided by the Data Capacitor allows one to collect large amount of information suitable for post processing, avoiding annoying in advance selection of subset of output data. The easy (from in- and out-campus) and open access (without application for defined limited computing time) allow both researchers and students to develop their applications in a significantly more comfortable manner and exploiting the specific hardware architecture provided by the cluster computing resources."

Highlight: Pre-Columbian ceramic vessels digitized and modeled

A collaboration between Anna VanderJagt of the CyberDH group and Jennifer St.Germain from the Glenn A. Black Laboratory at Indiana University has resulted in the digitizing and modeling valuable and rare artifacts – Mississippian ceramic effigy vessels from the collection of the Glenn A. Black Laboratory of Archaeology. Effigy vessels are one of the hallmarks of the pre-Columbian Mississippian era (~800CE – 1600CE). Humans, mammals, birds, fish, amphibians, and even mythological creatures were modeled in a variety of styles and vessel forms. Styles range from including one or more adornments that suggest certain features, such as a head and tail attached to opposite sides of a bowl rim, to entire bottles and bowls formed into the shape of a head or body. Because these vessels often contain asymmetrical features, they can be best appreciated and studied by viewing them from all angles and sides.

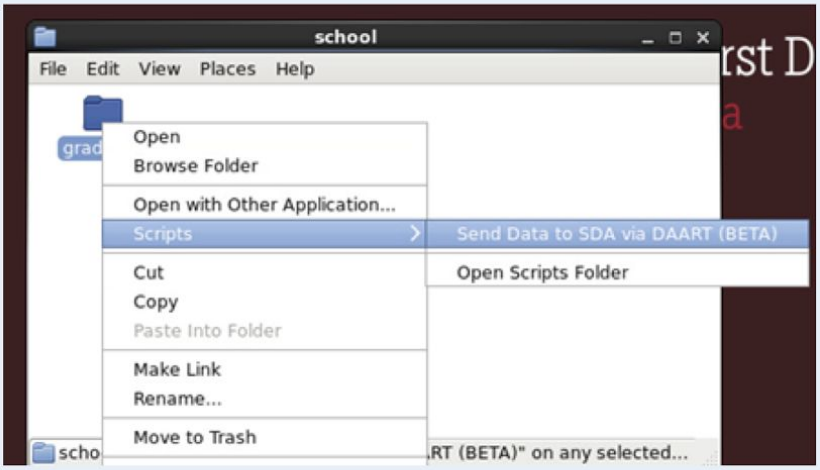


(Left) Jennifer St.Germain using static-object photogrammetry method on a human head effigy vessel from Glenn A. Black Laboratory collection. (Right) Female effigy bottle from Glenn A. Black Laboratory collection.

The creation of digital 3D models allows the public to dynamically view and engage with these artifacts in ways not otherwise possible when presented either in exhibit cases or in static 2D photographs. Researchers can also measure, analyze, and reconstruct these models in new and innovative ways. According to the team, "This project allowed us to get practical experience in the latest technologies in 3D modeling and gave us an opportunity to speak to many professionals in the IU community currently using these methods. We learned that every project has its specific goals and set of requirements, resulting in a wide variety of opinions on the best tools and technologies to use. In the end, we were able to provide the Glenn Black Lab with a set of resources and recommendations for selecting among these methods and successfully creating 3D models for research and exhibition." More details about this project including the methods and tools for creating and processing 3D models can be found here.

Highlight: Hit your data targets with DAART

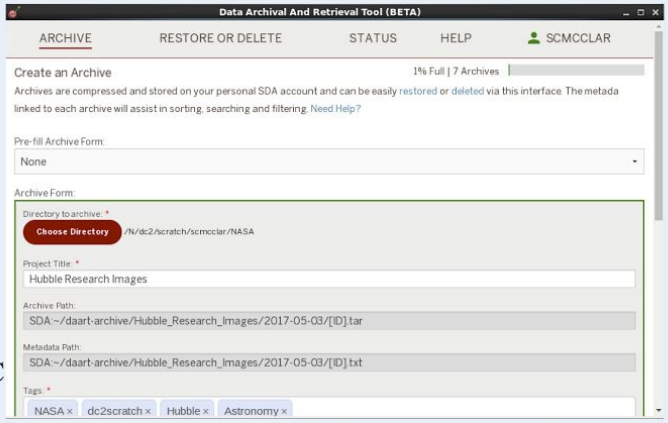
IU has robust supercomputing facilities and services that are free for IU's faculty, students, and staff to use. Once the computation and analyses are complete – IU offers long-term storage for data in the Scholarly Data Archive (SDA). Previously, SDA was largely accessed by command line and did not utilize metadata or "tags" to identify the data in a meaningful way. So, it was not straightforward to move data to the SDA and, once the data was on the SDA, it required a fair bit of searching through the files in the SDA in order find the desired file. To make moving the data and searching for it (later) easier, SciAPT and the SCA teams have developed a graphical tool called the Data Archive And Retrieval Tool (DAART).



Invoking DAART (Beta) from folder context menu

DAART enables HPC users to snapshot data from Data Capacitor II (DC II), or their home directory, or DCWAN II to the Scholarly Data Archive (SDA); and to restore previously stored datasets from the SDA. DAART's primary functions are:

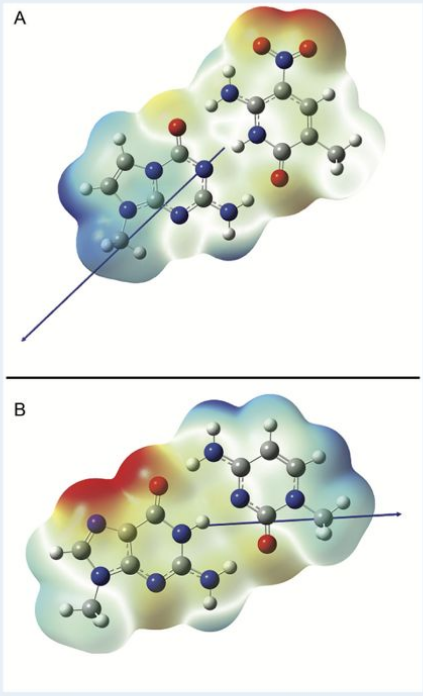
- Intuitive and seamless data movement between SDA and disk file systems (currently available within Karst Desktop Beta).
- Safe & secure archival of your valuable datasets; SDA replicates data between Bloomington & Indianapolis, in contrast to data getting purged on disk based filesystems.
- Freeing up critical disk space on DC II and DCWAN II for use by other users of IU supercomputing & storage CI.



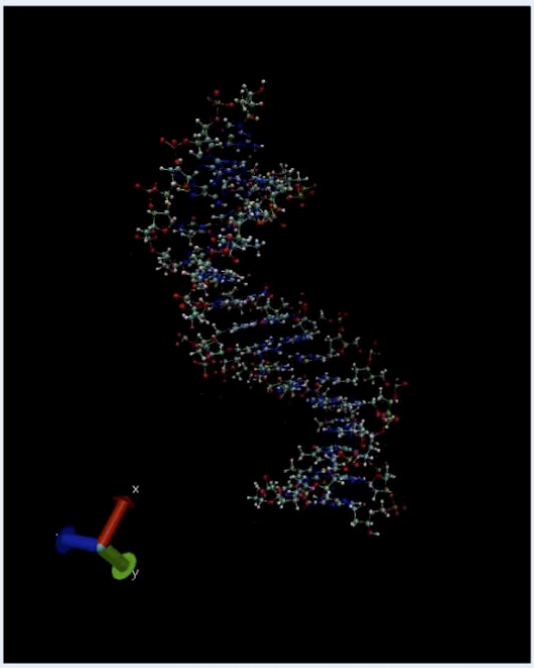
Screenshot of DAART options: Focused on archiving abilities.

Highlight: Simulating DNA to tackle tough medical challenges

Dr. Robert Molt Jr., a postdoctoral researcher in the Department of Chemistry and Chemical Biology at IUPUI and Department of Biochemistry and Molecular Biology at the Indiana University School of Medicine, has been conducting exciting new health-related research using IU's advanced cyberinfrastructure resources including Big Red II and Karst. Much of his work centers on the protein biochemistry mechanisms involved in cancer and generating artificial DNA nucleotides. According to Dr. Molt, "understanding how proteins work regarding cancer enables us to screen for inhibitors that can save lives." He also added that his work with artificial DNA allows for a deeper understanding of genetics and that his hope was to develop (stable) artificial genes to help people with genetic defects.



Dipole moments and the ESP for different nucleobase pairs



Still image of diffusing DNA helix

One of the toughest challenges that Dr. Molt and other scientist face is access to advanced computing facilities. Dr. Molt's work involves simulating DNA helices every $2x10^{-15}$ seconds of time for a total time of $5x10^{-5}$ seconds (i.e., 10 orders of magnitude in time!). According to Dr. Molt, "computing a dynamically stable trajectory over 10 orders of magnitude in time is only possible using the many GPUs of IU. I have been able to utilize ~150 GPUs at a time to achieve such calculations."

National and International Impact

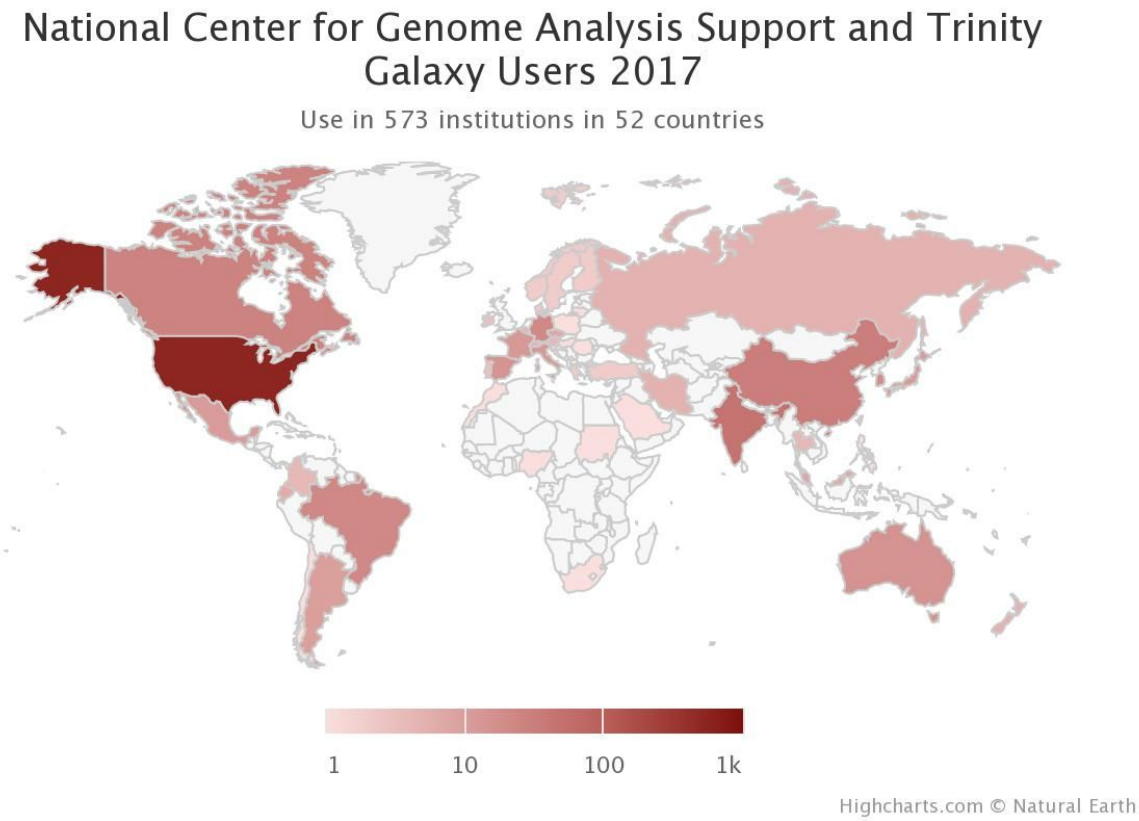
PTI Goal: Cultivate and enable creativity and innovation in science and scholarship by developing new innovations in cyberinfrastructure, informatics, and computer science

The PTI overall and particularly the PTI Centers, which operate as small and nimble R&D centers, lead the way in terms of development of innovations. Metrics describing new research and innovation achievements during FY 2017 are summarized below and described in more detail for each center.

Research and innovation productivity metrics for PTI from inception of PTL in 1999				
	Publications	Technical Presentation	Nobel Prize Awards Supported	Open Source Software Released
PTL & IUPUI 1999 - 2014*	1,796	1,426	3	178
PTI FY 2015	39	15	0	1
PTI FY 2016	34	55	0	8
PTI FY 2017	-9999	-9999	0	-9999
Total	-8,130	-8,503	3	

*Presentations from 2008 - 2014

The PTI is comprised of centers that have impacts that extend well-beyond the IU community. Below are highlights from the centers and projects which offer services to the nation and, in some case, to the world.



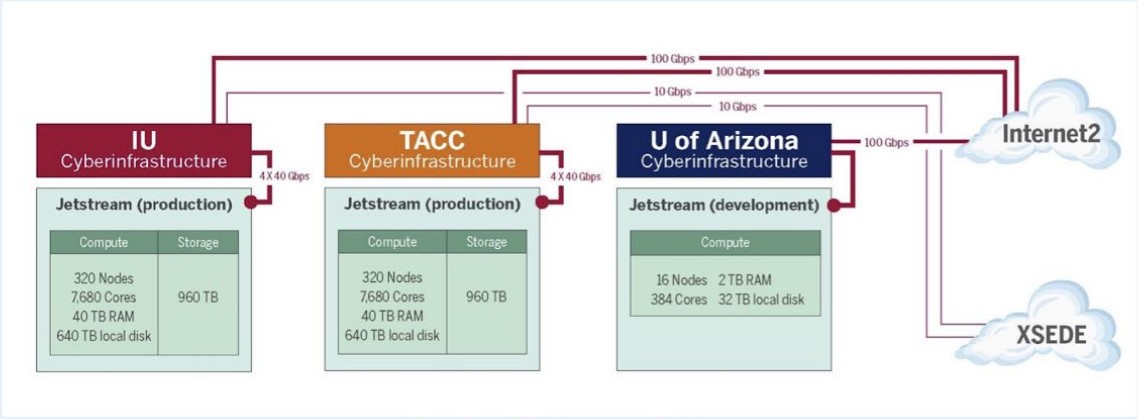
Robust Persistent Identification of Data (RPID) (NSF grant #1659310; PI Beth Plale):

This project focuses upon robust, persistent identification of data, which could greatly improve scientific discovery and reuse of datasets. Persistent identifiers of data (PIDs) enhance scientific discovery and are an important element of research data sharing. This project creates a testbed to evaluate new capabilities for persistent identifiers. The initial stage of the project will include four diverse repositories containing millions of PIDs, and the second stage will allow any NSF-eligible institution to use the testbed to evaluate their own work. The project would enhance research interoperability. Currently, there are several persistent identifier options for data (such as Digital Object Identifiers, Handles, the Archive Resource Key, and Uniform Resource Names). The existing environment is limited by multiple solutions, weak interoperability, and procedures translating PID to data object that are inconsistent. This project provides several new capabilities: a testbed to research new capabilities and interoperability for persistent identifiers, which builds upon Indiana University cyberinfrastructure and instances on Amazon Web Services; the ability to prototype and evaluate PID types, allowing study of the difficulties and advantages of relating types to one another across a distributed system; and approaches to mapping from PIDs to Canonical Text Services Uniform Resource Names (CTS URNs). Combining URNs with Handles should allow a precise CTS URN referencing capability, with the flexible resolution of the existing widely-used Handle System. The goal is to standardize the results of PID resolution, allowing various PID services to interoperate at a higher level.

Highlight: IU projects shine at 2017 Plant and Animal Genome Conference

The Plant and Animal Genome Conference is the largest Ag-Genomics meeting in the world: it brings together over 3,000 genetic scientists and researchers in plant and animal research and hosts 150 workshops and thousands of research presentations.

IU's National Center for Genomic Analysis Support (NCGAS) was focused on the Cenicafe collaboration with Cornell. The mission of Cenicafe is to improve the livelihoods and outcomes for coffee producers in Columbia, with particular focus on combatting coffee rust, a disease which threatens coffee growers and may percolate through the industry, drying up supplies for coffee consumers. NCGAS works with Cenicafe to analyze genomic information and host the results on genome browsers for easy interpretation. This collaboration is partially funded by grants (NSF #1444893) to support investigating transcriptomics of the Coffea arabica plant. Discussions with the Cenicafe group and Cornell enabled NCGAS to solidify goals for the next year, clarify the progress from the last year to all stakeholders, and strategize about how to proceed.

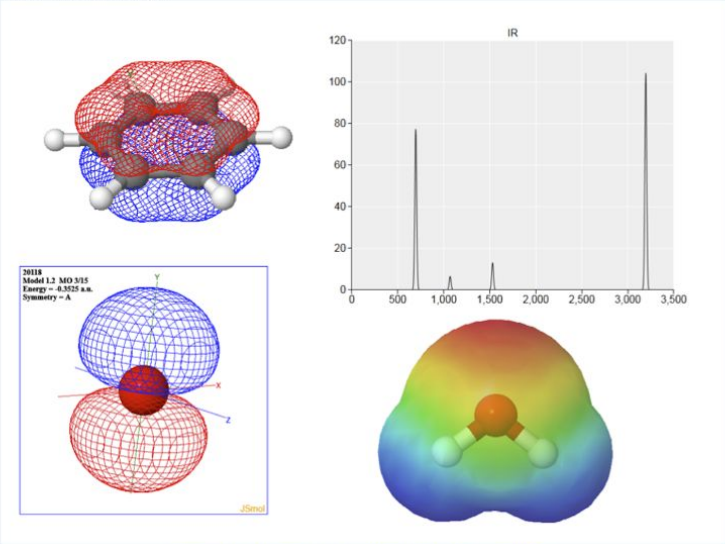


Overview of Jetstream architecture

IU was also represented by members of the Jetstream team. Jetstream (NSF #1445604) is a cloud computing system that provides on-demand services for life sciences research and education. The Jetstream team showed how different users can access the system: from laboratory scientists to informatics specialist to core facility staff and faculty.

Highlight: Chemistry students learn using Chemcompute and Jetstream

Mark Perri, Associate Professor in the Department of Chemistry at Sonoma State University, has been supporting thousands of students learning Chemistry across the nation using Jetstream. Perri runs a science gateway called Chemcompute that enables undergraduate students to calculate properties of molecules. A science gateway is a community-developed set of tools, applications, and data that are integrated via a web portal. According to Perri, there is a great need for science gateways like Chemcompute because, "Easy to use software to perform these types of calculations can cost thousands of dollars, not to mention the cost of hardware to run it on. There is high quality free software available, but it is not easy for undergraduates to use. Our science gateway provides a web-based portal that is easy for undergraduate chemistry students to use, eliminating the costly software. XSEDE resources, including Jetstream, provide the computational power to eliminate the need for hardware purchases." Perri added, "Undergraduates at small institutions might not have access to computational chemistry software at all without the use of our science gateway. We hope that we can spark an interest in these students to enable them to pursue postgraduate work in the field of computational chemistry."

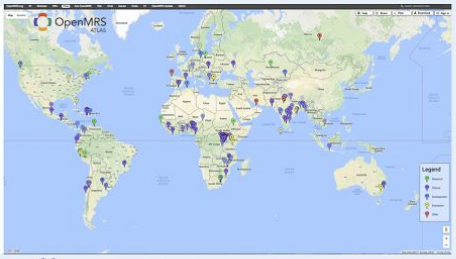


Examples of results produced by Chemcompute.

Since August 2016, Chemcompute served 21,200 jobs for 1,327 users consuming 33,000 Service Units (equivalent to vCPU hours). In order to run the Chemcompute science gateway, Perri has been utilizing resources provided by a local resource and Jetstream led by the Indiana University Pervasive Technology Institute (PTI). Jetstream is the 1st production cloud funded by the National Science Foundation (NSF grant #1445604). Jetstream allows researchers to create virtual machines on the remote resource that look and feel like their lab workstation or home machine, but are able to harness thousands of times the computing power. For Perri and Chemcompute, Jetstream is the perfect fit as far as resource providers go because Jetstream is 20 times faster relative to the local server and it provides both the web-hosting and computation resources required to run the gateway. Perri said that, "It has been difficult to find enough computational resources and web-hosting for all of our users. Typically one just has to worry about the needs of their own research group, not hundreds or thousands of users across the country. The types of computations our undergraduates run are also rather unique – many, many small jobs, rather than a few large computations." However, it appears to be smooth sailing for Chemcompute now with the aid of Jetstream!

Highlight: Write code, save lives

OpenMRS is a global project supported and led by Regenstrief Institute and Indiana University that provides an open-source electronic medical record system platform. Their core mission is to improve health care delivery in resource-constrained locations across the globe – spanning more than 80 countries. – with a compelling tag line "Write code. Save lives." OpenMRS community infrastructure operates on advanced cyberinfrastructure from Quarry Gateway Web Services Hosting System operated by Research Technologies. Soon, OpenMRS will transition to Jetstream infrastructure (NSF Award ACI-1445604). Jetstream is a National Science Foundation (NSF) funded cloud-based resource which provides access to computational resources to a variety of science and engineering research projects.



OpenMRS is in use around the world.

The OpenMRS community is a vibrant one that includes thousands of individuals from across the globe with clinicians, coders, and other volunteers around the world represented. OpenMRS was recently recognized when co-founders Burke Mamlin, MD and Paul Biondich, MD of Regenstrief Institute and Indiana University School of Medicine were awarded the 2016 Donald A.B. Lindberg Award for Innovation in Informatics by the American Medical Informatics Association (the world's largest international professional biomedical informatics association).

Photo taken at the OpenMRS Implementers Conference 2016 in Uganda.

More information about OpenMRS and volunteer opportunities can be found here.

Highlight: IU aids BigJack upgrade

As part of its Campus Bridging effort, the Extreme Science and Engineering Discovery Environment (XSEDE), the Capabilities and Resource Integration staff from Indiana University assisted in building a new HPC cluster at South Dakota State University in Brookings, South Dakota. Kevin Brandt, manager of the Networking and Research Computing team at SDSU, reached out to campus bridging staff asking for guidance on upgrading their existing cluster, BigJack, using the tools provided by XSEDE. The BigJack cluster currently serves about 100 active users on campus, enabling research in a variety of domain science areas.

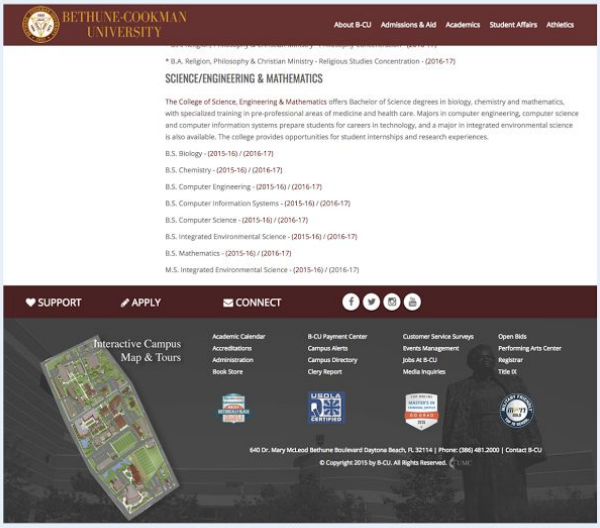
Over the course of a week, XCRI staff helped implement the XSEDE Compatible Basic Cluster toolkit on unused nodes, using the OpenHPC tools and XNIT (XSEDE National Integration Toolkit) repositories to provide modern cluster management and scientific software. The new cluster, named Campanile, will eventually integrate the nodes currently used in BigJack, resulting in a final size of 71 nodes, with 3360 TB of total RAM and 840 total cores. Tom Crowe, at Indiana University, gave additional help on the setup of Infiniband hardware, while Francesco Pontiggia (at Harvard as of May 2017) consulted based on his experience running an OpenHPC-based XCBC with authentication to local Active Directory resources.

From Left to Right: Maria Kalyvaki, HPC Domain Specialist, Chad Julius, Cluster System Support and Development, Brian Moore (front), HPC Domain Specialist, Kevin Brandt, Manager of Networking and Research Computing, and Eric Coulter, XCRI Engineer, in front of the Campanile and BigJack clusters. Photo credit: Nate Clapp

The new HPC cluster at SDSU will allow the research computing group to continue providing high quality service to their users, with a modern operating system and greater ease of transition to XSEDE resources. The new build ensures that SDSU admins are familiar with all the internals of their system, and will enable them to easily integrate all of their available hardware into the system, broadening their compute capabilities without additional spending.

Highlight: NCGAS provides HPC, training, and networking to Bethune-Cookman University students

During the Fall 2016 semester, 9 students were enrolled in a new bioinformatics course at Bethune-Cookman University, an HBCU (Historically Black Colleges and Universities) in Daytona Beach, Florida. In the course titled "Advanced Computing Resources in Biology", taught by Dr. Raphael D. Isokpehi, students learned the principles of command line genomics software and python programming as well as gained exposure to large-scale computing. This new bioinformatics course and two other courses: Biomolecular Technologies and Computational Genomics were developed at B-CU through National Science Foundation award #1435186 from the Historically Black Colleges and Universities - Undergraduate Program (HBCU-UP). The students also attended a guest lecture by staff from the National Center for Genomic Analysis Support (NCGAS) at Indiana University. Four students met with NCGAS at the 2016 Supercomputing Conference in Salt Lake City, UT. They were able to inquire about the bioinformatics career path from industry professionals and gain insight into what to expect when moving from biology heavy training to computational-based science.

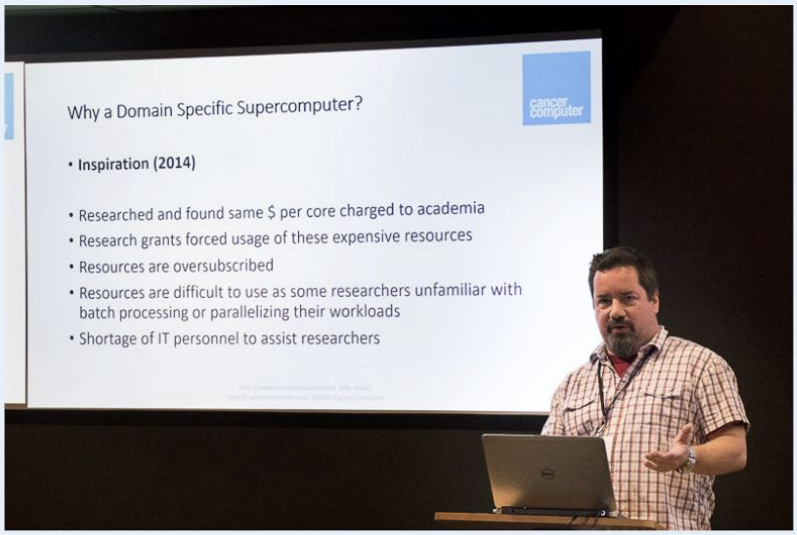


Screenshot from Bethune-Cookman University's listing of degree programs for science and engineering

Bioinformatics is a fast growing field, but training requires university level training in cluster computing and genomics software. "The collaboration with NCGAS is enabling biology students at Bethune-Cookman University to access computing resources for working effectively with large-scale biological data" says Dr. Isokpehi. Students are also acquiring the expertise to compete for internships and fellowships in pursuit of a career in biology or bioinformatics. Providing science and engineering cyberinfrastructure and training to Bethune-Cookman University and other institutions helps expand the curriculum available without the overhead cost of institution level scientific computing infrastructure. In person interaction with students and new users at conferences solidifies relationships and increases NCGAS's presence in the community. As the course was successful in providing previously inaccessible resources for training students, NCGAS will continue working with Bethune-Cookman in this capacity this Spring 2017 semester with a cohort of 18 students enrolled in a third-year introductory bioinformatics course.

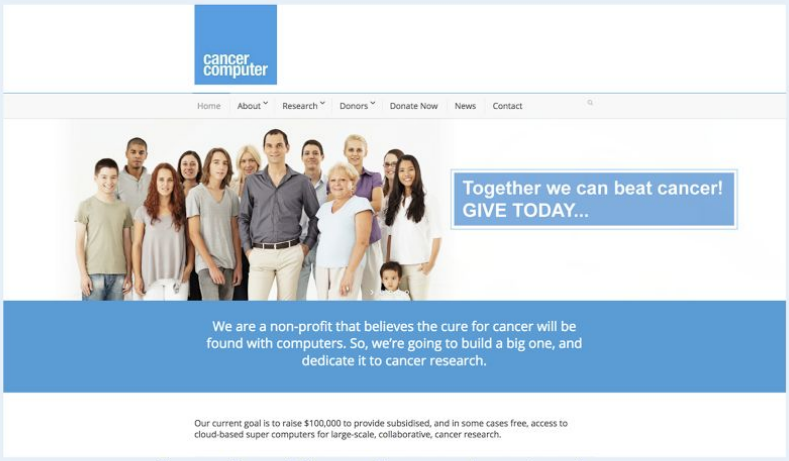
Highlight: OSG and Cancer Computer partnering for the cure

Cancer Computer recently partnered with Open Science Grid to make available 1,157,122 core hours per year for research at Indiana University School of Medicine (SPLinter project) and Harvard School of Medicine (Structural Biology Grid Science Portal).



Roy Chartier presenting about Cancer Computer at the Open Science Grid All Hands Meeting held March 6-10, 2017 at the San Diego Supercomputer Center in La Jolla, CA (photo by Kyle Gross)

In a recent interview with the OSG, Roy Chartier, founder and chief technology officer of the organization said, "The mission of Cancer Computer is to help accelerate the cure for cancer. We make computing resources, made possible through donations and corporate goodwill, available to research projects, such as those served by the Open Science Grid. That's why we partnered with the OSG." He added, "We provide resources directly to researchers, who either need more resources than are available to them departmentally, or need more resources than they otherwise might be able to afford on a commercial cloud platform. In many cases, we will allow researchers to use our platform at no charge." A full version of the original story can be found here.



Screenshot of Cancer Computer's main webpage

In order to further facilitate the exchange of ideas through conferences and workshops, the PTI also aids in the management of conferences and workshops held at other locations, described in the table below.

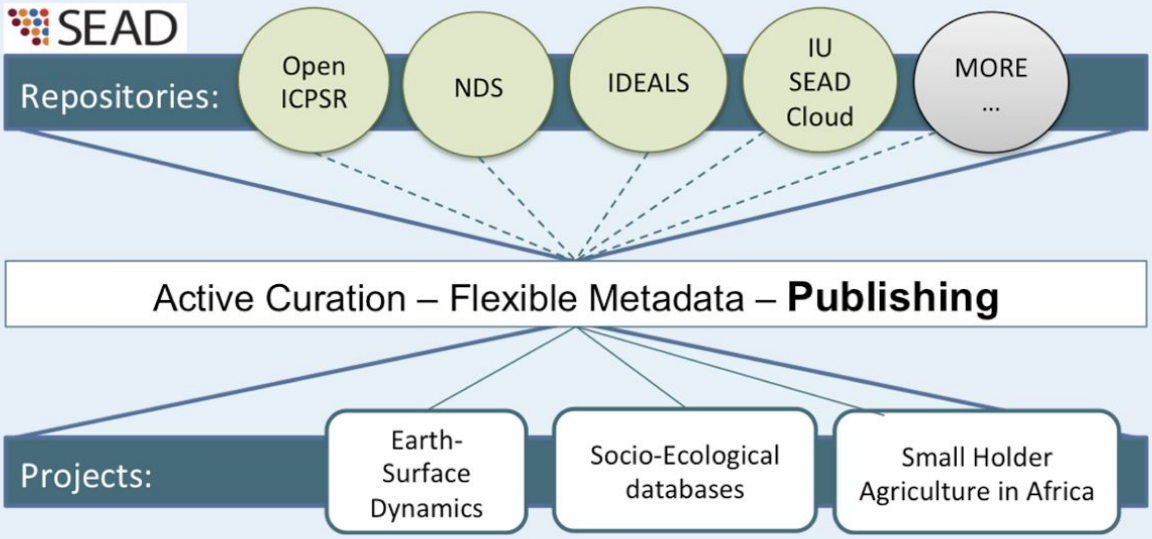
Academic conferences and workshops organized by IU and other collaborators, held at locations other than IU				
Conference	Topic	IU attendees (total)	Faculty attendees from outside IU	Non-IU attendees (total)
Environmental Genomics Workshop	Genomics; hosted at Salisbury Cove, ME	0	15	18
Workshop on High-Performance & Distributed Cyberinfrastructure for Polar Sciences: Applications, Requirements and Opportunities	NSF-sponsored workshop on use of distributed cyberinfrastructure for polar research applications, identifying strategies for supporting polar research in the field and after data collection			
Open Science Grid All Hands Meetings	Distributed High Throughput Computing	10	60	60
SPXXL Winter Workshop (Feb 2017)	SPXXL is a user group for organizations that have large installations of IBM or Lenovo equipment. The focus of this meeting was engaging with new vendors and partners in technical NDA discussions (Lenovo, Intel, NVIDIA, and Mellanox). IU staff led all aspects of accommodation and meeting venue planning and contract negotiations.			
Cray User Group Conference (May 2017)	The Cray User Group is an independent, international corporation of member organizations that own Cray Inc. computer systems. Founded in 1978, CUG was established to facilitate collaboration and information exchange in the high-performance computing (HPC) community. IU staff led the overall event organization as chair of the CUG board.	7	15	203
SPXXL Summer Workshop 2016	SPXXL is a user group for organizations that have large installations of IBM or Lenovo equipment. The focus of this meeting was engaging with new vendors and partners in technical NDA discussions (Lenovo, Intel, NVIDIA, and Mellanox). IU staff led assisted in accommodation and meeting venue planning and organized the member site			

	program contributions.			
XSEDE16 (July 2016)	The XSEDE conference series is a venue for researchers using XSEDE supercomputing and advanced research infrastructure, XSEDE support and operations staff, XSEDE outreach and campus representatives, and other interested groups to present research work, participate in Birds of a Feather and panel sessions, and organize tutorials.			
IEEE/ACM Supercomputing Conference 2016 (SC16; Nov 2016)	IU hosted a booth with interactive demos on a variety of projects currently underway in UITS Research Technologies.			
HUF 2017	High Performance Storage System (HPSS) Users Forum is a meeting hosted annually by an HPSS customer site to engage customers, HPSS developers and other relevant HPSS staff for current issues and future plans discussions. IU staff were part of the Steering Committee and Technical Program Committee and annually lead the Burning Issues session (for tracking progress on customer reported issues).			
Total		17	90	281

Highlight: D2I releases SEAD 2.0

The Data to Insight Center (D2I) at IU has released SEAD 2.0 It has an improved project space, customized metadata, streamlined publishing with several repository choices, and better tracking of user publication requests. In particular, the newly designed SEAD publishing pipeline, now consists of:

- **Curbee**: a suite of microservices that is applied to research objects to enhance (curate) them for publication.
- **SEAD Matchmaker**: a recommendation engine that identifies most appropriate repositories for publication and publishes research objects.
- **PDT (People, Data, Things)**: a repository of profiles about the people, data, and repositories that are used in SEAD



The IU SEAD Cloud, another component of the 2.0 suite of tools, is a reusable, thin preservation and repository layer over an HPC storage system. It is deployed at IU over an HPSS replicated tape store. More details about SEAD@IU and people involved can be found here. For a list of new features, click here.

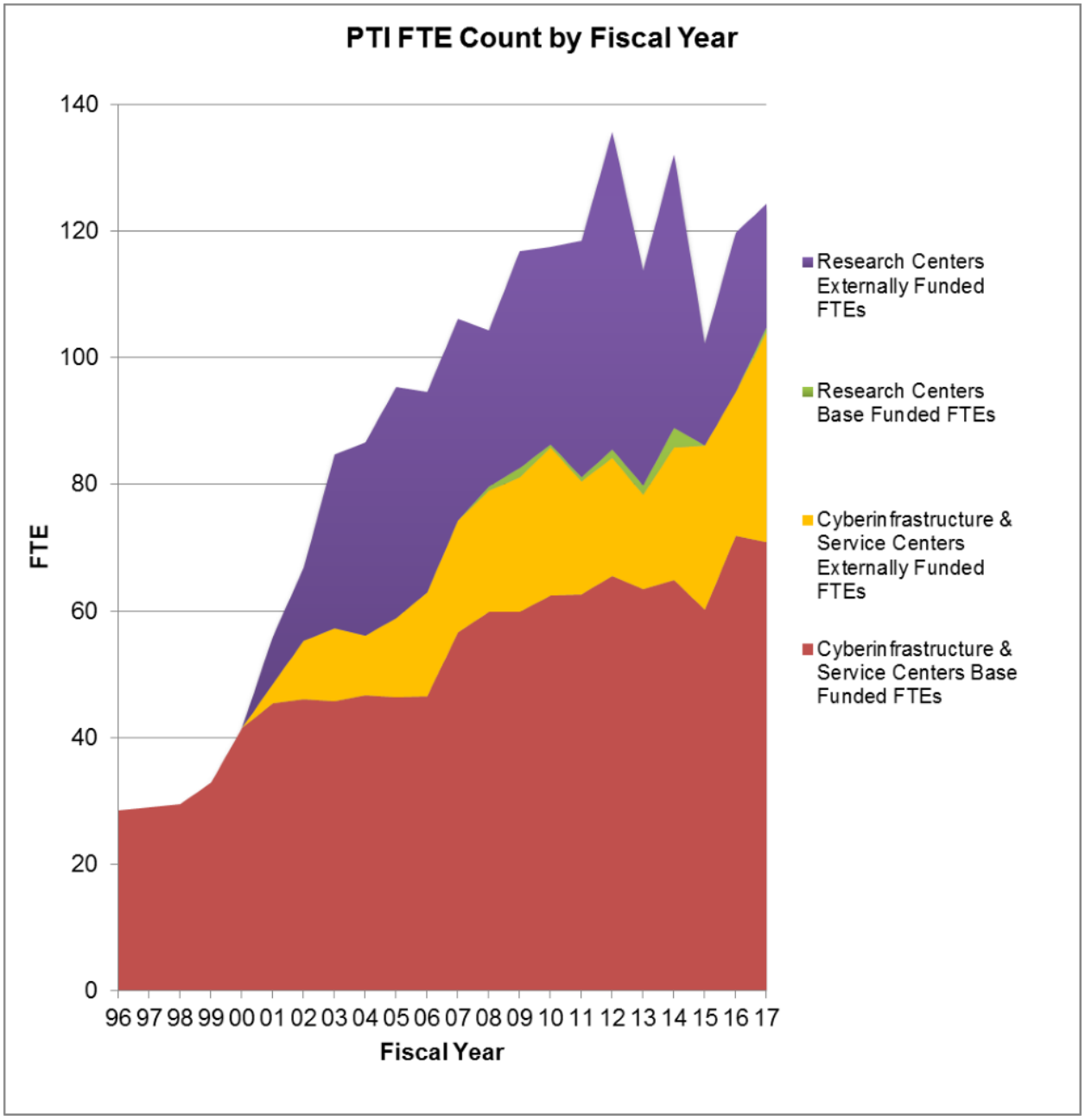
SEAD is funded in part by the National Science Foundation under its DataNet program, grant # 0940824. The Data To Insight Center carries out foundational research in the social and technical challenges of data use in research and scholarship and develops innovative tooling and cyberinfrastructure to advance science and scholarship.

Economic Development Impact

PTI Goal: Impact the economic health and quality of life in Indiana – creating new jobs, nurturing new businesses

The mission of the Indiana University Pervasive Technology Institute (PTI) is to improve the quality of life in the state of Indiana and the world through novel research and innovation and service delivery in the broad domain of information technology and informatics.

One of the most important ways in which PTI enhances the economic health and quality of life in Indiana is by being successful in the highly competitive process of winning federal grants and contracts, and creating new high quality jobs in Indiana as part of IU. Such jobs have an average salary greater than the overall average in Indiana, add to tax roles, and often bring highly qualified professionals from other states or nations to Indiana, where they very often settle down and stay their entire careers. Below is the number of employees (one employee = one FTE) by type of funding.



PTI Goal: Support the development of a 21st century workforce within the State of Indiana

PTI supports the development of a 21st century workforce in a variety of ways. One of the primary ways is through education offered by faculty with a PTI affiliation. As stated earlier, course delivery activities result in academic credits at IU, and degree-granting programs are offered primarily through PTI-affiliated centers that are also subunits of the IU School of Informatics and Computing (D2I and DSC in particular). PTI provides a number of services and facilities that support the development of a 21st century workforce in Indiana by facilitating success and innovation by current IU students, and by

spurring interest in STEM disciplines (Science, Technology, Engineering, and Mathematics) through K-12 outreach programs.

In 1955 astronomy professor Marshall C. Wrubel was appointed as the first director of the IU Research Computing Center. One of the first things Dr. Wrubel did was determine that his own graduate students, and the graduate students of others, were as worthy of using IU's electronic research computer as any IU faculty member. Thus began a 60-year (and counting) commitment to support for IU student achievement by the research computing center and its successors. Current critical services offered to undergraduate and graduate students of IU include access to software (analytical, geographic information system, mathematical, and statistical) and access to and use of supercomputers. Below are highlight of the services available to students and their usage.

- **Access to a large suite of software.** IUanyWare is a client virtualization service available to Indiana University students, faculty, and staff. With IUanyWare, you can use a web browser or mobile app to run certain IU-licensed software applications without having to install them on your computer or mobile device. In FY 2017, more than 48,000 undergraduate students and 7,000 graduate students accessed software through IUanyWare.
- **24/7/365 use of the entire suite of advanced visualization technologies.** Available technologies include ultra-resolution IQ-Walls, interactive and collaborative IQ-Tables and IQ-Tilts, 3D scanning equipment, spherical displays, and a variety of interfaces and displays that support virtual and mixed reality. Of particular note, IU Bloomington has seen increased student use of IQ-Walls in public spaces (e.g., the main library). IUPUI has seen tremendous uptake from students because of their close partnerships with Informatics. Informatics students met and utilized the provided visualization technology nearly daily, and often on weekends and evenings. A loaner program for portable equipment (the most popular being the Oculus Rift Development Kits) affords additional opportunities for independent learning and exploration.
- **Use of the GitHub Enterprise Service.** IU is the first academic institution to provide the GitHub Enterprise Service for distributed source code control, software development, and collaboration. The git distributed version control system allows considerable flexibility and ease of collaboration in a networked environment. This service has seen considerable uptake across the university, not only from the School of Informatics and Computing, but also in numerous academic units as well as in administration. Currently github.iu.edu hosts 11,882 repositories owned by 3,054 users, supporting student projects and assignments, departmental web services, faculty collaborations, and enterprise software development.

Student Users by Resource									
	Undergraduate students			Graduate students			Total		
Type of system	FY 2015	FY 2016	FY 2017	FY 2015	FY 2016	FY 2017	FY 2015	FY 2016	FY 2017
IUanyWare	--	48,745	46,829	--	> 2,000	7,038	--	48,829	64,671
Supercomputers & computational	36	23	24	558	331	538	605	354	562

systems									
Advanced storage systems	282	58	87	1,157	484	670	1,477	542	757
Advanced visualization systems	144	168	9	73	70	108	217	238	117
RT GitHub code repository	1,640	1,638	2,300	1,154	1,189	1,511	2,794	2,827	3,811
Total	2,102	50,632	49,249	2,942	2,074	9,865	5,093	52,790	69,918

Employment, education, and practical experience for IU students

Two PTI research centers are part of the School of Informatics and Computing, headed up by leading IU faculty members. Such centers have the education of undergraduate and graduate students as a core part of their mission. The other PTI centers also provide employment and practical research experience for undergraduate and graduate students.

During the first 15 years of PTL and PTI activities, PTI provided research experiences to 203 students, including: 66 PhD students and post-doctoral fellows, 116 masters students, and 21 undergraduate students. One particular note on graduate student success is that Richard Knepper, manager of Campus Bridging and Research Infrastructure, completed his PhD in the School of Informatics and Computing.

Overview of Student Projects			
Student	**Institution**	**Classification**	**Project Description**
Anna VanderJagt	Indiana University	M. Information Science	Photogrammetry and web-creation and design for various departments at IU
Gagan Deep	Indiana University	M.S. Computer Science	Cluster viewer for Astronomy Department at IU
Megan Eller	Indiana University	M. Information Science	Information architecture; web design and accesibility for AVL

Anna VanderJagt, graduate student intern with CyberDH group in RT

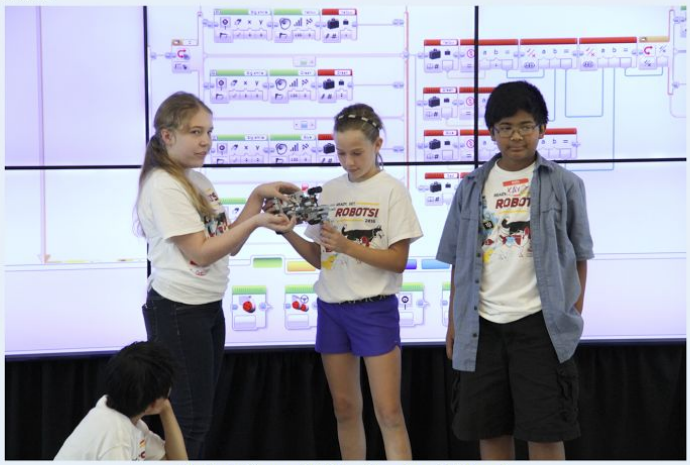
Education and outreach activities that interest and inspire the scientists and technologists of the future

Of course, people don't generally wake up one morning when they are 16 or 17 and fire up a supercomputer. The PTI supports a number of prominent outreach events that are community-focused. Highlights of these education and outreach events are below.

- **Jim Holland Research Initiative in STEM Education (RISE).** The Jim Holland Research Initiative in STEM Education (RISE), run by the IU Department of Biology is the third-year component to the Holland SEP and SSRP. It is a two-week residential research camp designed to provide STEM career and college training for rising high school seniors (students currently in grade 11). RISE scholars attend research-intensive lectures and participate in hands-on activities delivered by IU faculty and staff members. In FY 2017, PTI hosted one day of the activities and had participation from staff from Jetstream, CACR, CyberDH, AVL, NCGAS, CREST (Center for Research in Extreme Scale Technologies) and CESG.

Highlight: PTI's (2016) Ready, Set, Robots! program is as fun and awesome as ever

Encouraging youth to enter a field of study in science, technology, engineering, or mathematics (STEM) continues to be a focus of the outreach and education component of IU PTI. The Ready, Set, Robots! (RSR) program, now in the 9th year, is one of the many programs IU offers in the area of K-12 STEM outreach.



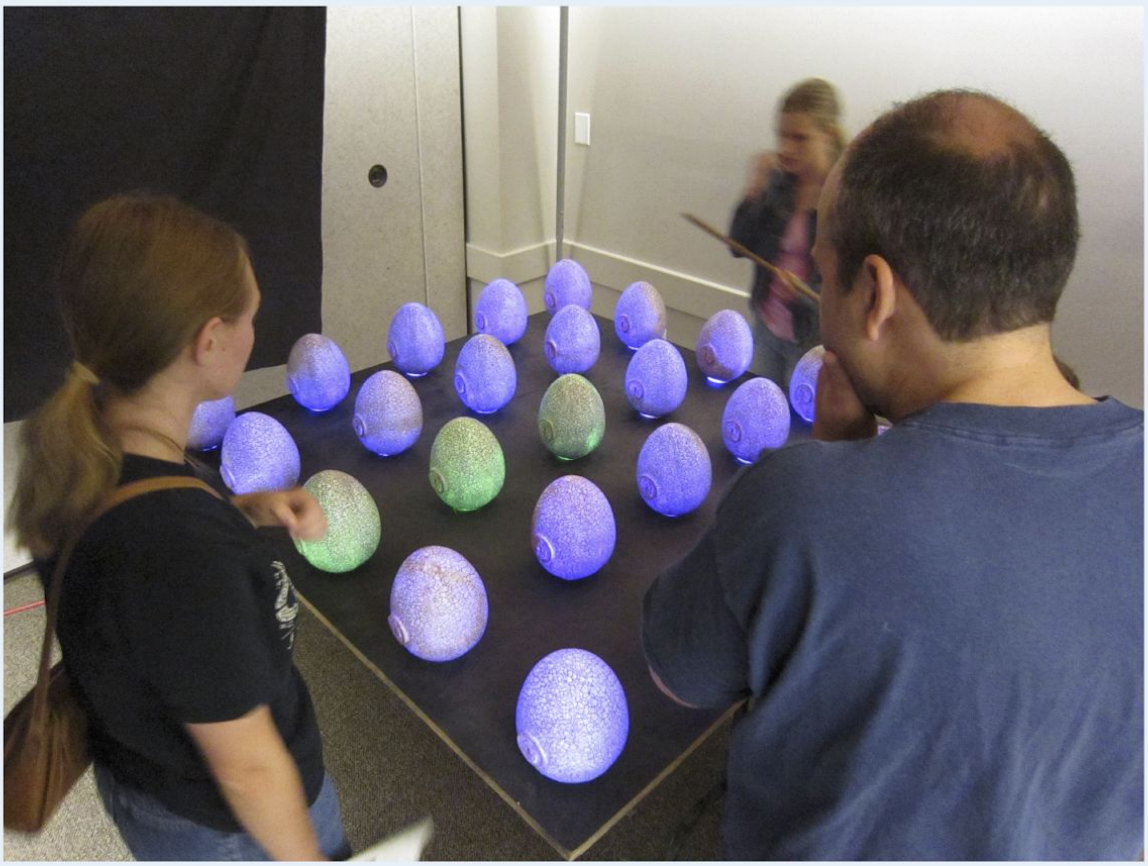
Robot Camp 2016 - Session I, Day 2

The 2016 RSR two-day camps were held at IU Bloomington and introduced youth to computer programming using Lego® Mindstorm® robots. In addition, campers dabbled in public speaking, experienced advanced visualization through use of the IQ-Wall in the Cyberinfrastructure Building for final presentations, and toured the Data Center where high performance computers such as IU's Big Red II an Mason are located. 48 students attended the 2016 beginner camp where they completed their "mission to Mars", and an estimated 100 friends and family attended the Robot Grand Challenge where the campers present the results of their newly-learned programming knowledge. This year, we hosted the 2nd Advanced RSR Camp; 16 students learned how to "talk" in binary and used some of the advanced features of the program to successfully complete their search and rescue mission.

Robot Camp 2016 – Session III, Advanced Camp

Highlight: Makevention 2016

Makevention is where maker groups in Bloomington area communities come together to share the do-it-yourself spirit with each other and the community. These makers encompass a broad range of fields, including tech enthusiasts, artists, educators, crafters, hobbyists, and tinkerers. Sponsored in part by IU Research Technologies – Pervasive Technology Institute, Makevention (August 2016) attracted 1200 people with 25 groups exhibiting. Bloominglabs, a local hackerspace, helped organize the event. Jenett Tillotson and Nathan Heald from RT play a leading role in supporting Bloominglabs and organizing events such as Makevention.



Lights out! One of the activities at Makevention 2016. Photo by Nathan Heald.

This event drew together the local community of makers of all ages. Attendees sparred with the Shire of Mynydd Seren from the Society for Creative Anachronisms, explored the myriad of projects from Bloominglabs, learned about locks from the Fraternal Order of Lock picking Sport and Columbus Key, built (and destroyed!) cup towers with Wonderlab, solved mechanical puzzles with the Lilly Library, folded gigantic origami figures with Discardia, browsed hand-made items from locals, mixed additives into soap with Soapy Soap Company, played life-size foosball games, viewed the Trashion Refashion exhibit, and more! Fun was had by all while also learning about science, engineering, or building through hands-on activities.

PTI Centers

Center for Applied Cybersecurity Research

Dear friends of the Center for Applied Cybersecurity Research

Cybersecurity was constantly in the news this last year, touching on elections, medical devices, smartphones, critical infrastructure, and our day-to-day privacy. CACR is tackling this challenge both directly, working with organizations across the Nation to improve their cybersecurity posture, and broadly through applied research and outreach. Highlights of our impacts from the past year are:

- CACR continues its leadership role in the National Science Foundation's Cybersecurity Center of Excellence, helping secure more than $7 billion dollars of research funded by the NSF.
- Continued collaboration with Crane Naval Surface Warfare Center. As two of Indiana's leading institutions in addressing cybersecurity, we exchange personnel and collaborate, bringing both organizations strengths to bear on some of the Nation's toughest cybersecurity challenges.
- As one of the lead institutions in the Department of Homeland Security Software Assurance Marketplace (SWAMP) we are working to improve the software that serves as the foundation of our phones, medical devices, and cars. In late 2016, the SWAMP launched a new version of its continuous assurance technologies, called "SWAMP-in-a-Box" (SiB), that allows the software assurance community to deploy on premise instances of the SWAMP.
- Leading a new $1 million award from the National Science Foundation, Institute to assure scientific workflows and data.
- Supporting colleagues at Indiana University's School of Informatics and Computing, Kelley School of Business, and Maurer School of Law in launching IU's Cybersecurity Master's program.
- Welcoming of three new Fellows: Robert Cowles of BrightLite Information Security and former CISO at SLAC National Accelerator Laboratory, Rachel Dorkery from the Maurer School of Law and Robert Templeman, Chief Cybersecurity Engineer from NSWCCD.
- Providing education and workforce development through training, talks, and workshops in Indiana and throughout the Nation.

These examples, along with with other activities described in this report, are the foundation of the pride CACR has in holistically tackling our Nation's cybersecurity challenges.

About the Indiana University Center for Applied Cybersecurity Research

CACR is distinctive in addressing cybersecurity from a comprehensive, multi-disciplinary perspective, by drawing on Indiana University's wide range of scholarly expertise in computer science, informatics, accounting and information systems, criminal justice, law, organizational behavior, and public policy, as well as the extensive practical cybersecurity experience of its operational units. Founded in 2003, CACR is a research center affiliated with the Indiana University Pervasive Technology Institute and a member of the Indiana University cybersecurity community, which includes the Maurer School of Law, the Kelley School of Business, the School of Informatics and Computing, REN-ISAC, the University Information Policy Office, and the University Information Security Office.

CACR Mission & Vision

CACR is Indiana University's flagship center for cybersecurity, serving as an integrator for research across IT's different schools and organizations. Its mission is to empower people with the knowledge and skills they need to manage cybersecurity risks. It does this through an applied research cycle of undertaking cybersecurity operations, particularly in unconventional settings, and learning from those endeavors as well as the research and lessons of others. It then applies those improved outcomes in its own work and disseminates them broadly through education, training and engagement. CACR is devoted to interdisciplinary cybersecurity and tackling challenges holistically across technical, policy, and social factors.

CACR'S Major Initiatives for 2016-17

CACR is proud of the following impacts its initiatives had for the Nation.

CACR Collaborates with NSWCCD on Cybersecurity

In July 2016, CACR, represented by Vice President for IT and CIO Brad Wheeler, and Naval Surface Warfare Center Crane Division (NSWCCD), represented by then-Crane Commanding Officer Captain JT Elder signed a 2 year Cooperative Research and Development Agreement (CRADA) to collaborate on cybersecurity as it applies to some of our nation's most critical challenges. Over the year since the start of the collaboration, the collaborative work has been presented over a dozen times to DoD senior leadership, including at the Pentagon. NSWCCD's lead for the collaboration, Rob Templeman, was promoted to Chief Engineer and designated a CACR Senior Fellow. CACR staff Jackson and Russell have been invited to apply for prestigious temporary faculty positions at NSWCCD.

CACR Provides Cybersecurity Expertise to Open Science Grid.

https://www.opensciencegrid.org/

The Open Science Grid (OSG) is a nationwide facility and infrastructure enabling large-scale high-throughput computing for science. In FY16, CACR joined the collaboration, working alongside the IU's Grid Operations Center, providing cybersecurity leadership for OSG. In FY16, CACR's focus was improving security services through increased automation, fixing weaknesses in traceability for certificate-free jobs, beginning deployment of secure hardware tokens for cryptographic keys needed to automate host certificate signing and deployment, and growth in areas such as software assurance and security program management.

Scientific Workflow Integrity with Pegasus
https://cacr.iu.edu/projects/swip/
CACR is leading a new three-year project funded by a $1 million grant from the National Science Foundation. The Scientific Workflow Integrity with Pegasus (SWIP) project, in collaboration with Dr. Steven Myers at the Indiana University School of Informatics, the Renaissance Computing Institute, and University of Southern California Information Sciences Institute, will improve the security and integrity of scientific data by integrating cryptographic integrity checking and provenance information into the Pegasus workflow management system.

CACR Collaborates with ICEI
https://icei.org/
ICEI formed to address the reality that too much of the internet's infrastructure is supported by too few people. The organization's mission is to support the development and stewardship of reliable, secure, and open source internet infrastructure software. In particular, ICEI focuses on the software underlying the internet's critical functions, modernizing it and future-proofing it by providing developer resources and expertise.

ICEI's deep technical connections and experience securing internet software, combined with CACR's cybersecurity expertise, make for a powerful collaboration to bolster internet security. The two organizations will collaborate on making the internet more secure and more reliable, and in raising financial support through federal grants and other public or private sector support to sustain these efforts. Such funding is crucial to addressing the underlying causes of internet software vulnerabilities: lack of support, professional development, and a maintenance network.

"The skill and manpower crisis in internet infrastructure software is formidable," said Andrew Kirch, chairman of the ICEI board. "CACR's support is invaluable in overcoming these challenges to preserve a reliable, secure, and open internet for everyone. We're happy to be working with CACR."

ICEI and CACR have a rich history of working together. They previously collaborated on a new version of the Network Time Protocol (NTP). Like too much of the internet's infrastructure, NTP was out-of-date and increasingly vulnerable. One part-time person supported the critical time-keeping software, and had

lost the root passwords to the machine where the source code was maintained (so it went years without security updates).

Software Assurance Marketplace (SWAMP)
continuousassurance.org
CACR is one of four institutions leading the SWAMP project, which was launched in 2012 by the Department of Homeland Security-Science & Technology Directorate to advance the effectiveness of software assurance technologies and to expand their adoption by software developers. The SWAMP's provides an integrated, one-stop environment for developers to analyze their code with multiple tools and offering a unified assessment results viewer.

In late 2016, the SWAMP launched a new version of its continuous assurance technologies that allows the software assurance community to deploy on premise instances of the SWAMP. This version, called "SWAMP-in-a-Box" (SiB), augments the services provided by the public SWAMP facility that has been in operation for the past three years. It is a free, self-contained version, including 15 open source tools, that can be installed on local servers or individual computers, addressing the need of organizations that must or prefer to keep their software assurance activities on premise. This marks a major step in the SWAMP's now four-year effort to bring continuous software assurance capabilities to mainstream code developers.

The Information Security Practice Principles
https://cacr.iu.edu/principles/
CACR developed the Information Security Practice Principles based on its own experiences in providing cybersecurity cybersecurity leadership to communities. Over the last year they have become a flagship CACR offering in their own right to the security community at large. They are the foundation of our collaborative work with NSWCCD, the subject of briefings to senior leadership at Pentagon and DoD leadership, and presentations by CACR staff at O'Reilly OSCON.

NSF Cybersecurity Center of Excellence
trustedci.org

Led by CACR, the NSF Cybersecurity Center of Excellence (CCoE), completed its first 18 months as a clear success. The CCoE undertook one-on-one collaborations with nine different NSF project (The US Antarctic Program, Gemini Observatory, WildBook/IBEIS, Array of Things, SciGaP, HUBzero, OSiRIS, the TransPAC IRNC network, and IGO) to address their cybersecurity challenges. It organized and hosting the NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure in Arlington, VA, with 100 attendees. Additionally the NSF Summit provided 15 highly-rated training sessions topics to

over 130 professionals on cybersecurity topics including identity management, log analysis, and secure coding.

2016 NSF Summit attendees

The CCoE undertook keep partnerships to advance cybersecurity. With the Department of Energy's Energy Science Network, it developed the Open Science Cyber Risk Profile to guide science projects in better understanding and addressing risks to valuable scientific assets. In collaboration with NSF's newly launched $15m Science Gateway Community Institute (SGCI), the CCoE is tackling science gateway security, a key means of using the world wide web to make science broadly accessible.

Infrastructure for Privacy-assured CompuTations
https://cacr.iu.edu/news/2017/RENCI-and-CACR-partner-on-NSF-project.php
CACR is contributing its cybersecurity expertise to a new three-year, $3 million project, funded by the National Science Foundation. The Infrastructure for Privacy-assured CompuTations (ImPACT) project, led by the Renaissance Computing Institute, will allow researchers to focus more fully on science by building a technology infrastructure that supports best practices in moving data, managing data, ensuring security and preserving privacy.

Array of Things
arrayofthings.github.io

The Array of Things (AoT) is an urban sensing project, a network of interactive, modular sensor boxes that will be installed around Chicago to collect real-time data on the city's environment, infrastructure, and activity for research and public use. AoT will essentially serve as a "fitness tracker" for the city, measuring factors that impact livability in Chicago such as climate, air quality and noise. CACR is collaborating with the AoT team to lead their efforts related to cybersecurity and privacy.

Educating the Nation on Cybersecurity
IU Cybersecurity Master's Program
https://cybersecurityprograms.indiana.edu/

CACR supported the Indiana University's School of Informatics and Computing, Kelley School of Business, and Maurer School of Law in launching Indiana University Cybersecurity Master's program. This two-year degree provides the foundation for our workforce to address cyber threats to national and international security through holistic training in privacy, intellectual property, and information and systems security.

CACR Speaker Series draws experts from across the country.

CACR sponsors a bi-monthly Security Seminar featuring internal and external experts who present their current research and real-world experiences to IU faculty, staff, and students. These presentations are offered at IUPUI and IU Kokomo via live stream. The 2016-2017 Speaker Series, attended by 266 individuals, featured the following experts speaking on the listed topics:

Patrick Traynor
University Of Florida
Who do I think you are?

Mohammad Khan
University of Connecticut
Understanding & Altering Users' Motivation to Follow Computer Security Advice

Jessica Staddon
North Carolina State
Privacy Incidents, News & News About Incidents

Nate Cardozo
Electronic Frontier Foundation
Encryption and the Law

Stephanie Pell
West Point's Army Cyber Institute
Broken

Brad Wheeler
Indiana University VP for IT & CIO
Its worse than you think… and what to do about it.

Jackie Kerr
Lawrence Livermore National Laboratory
Authoritarian Soft Power?

LTC Ernest Wong
West Point
The Next Big Idea

Ryan Gagnon
New York Army National Guard
The Next Big Idea

Media Appearances

To help educate the public and advance the National discourse on cybersecurity, CACR works with the press press to provide background and clarity regarding cybersecurity. Some highlights of our media appears include:

- Being highlighted in the Coalition for Academic Scientific Computation 2017 Brochure: http://casc.org/page/Brochures
- Susan Sons on maintaining and securing the internet's infrastructure: https://www.oreilly.com/ideas/susan-sons-on-maintaining-and-securing-the-internets-infrastructure
- Susan Sons talks about Trump-era FCC regulations for your internet browsing history: https://www.youtube.com/watch?v=aExb49k5yuE
- Von Welch talks to the BBC about Smart City security: http://www.bbc.com/news/business-36854293
- Susan Sons talks to the Observer about Open Source Software security: http://observer.com/2016/11/open-source-too-big-to-fail/

NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure

Through the Center's leadership of the Center for Trustworthy Scientific Cyberinfrastructure (CTSC), CACR executed the NSF Cybersecurity Summits for Large Facilities and Cyberinfrastructure in 2016 as it has for the past four years. One hundred members of the community attended for three days of training, presentations and discussions on the cybersecurity challenges faced by NSF scientific research.

CACR Cybersecurity Summit

CACR has been bringing together leading visionaries in the area of applied cybersecurity technology, education, and policy in an annual Cybersecurity Summit since 2010. During this one-day event, attendees discuss the proper balance of public needs, homeland security concerns, and individual privacy rights. The 2016 CACR Cybersecurity Summit focused on *Privacy vs. Security*, and featured keynote Drew Minnick from AccessNow and Kevin Branzetti from the New York County District Attorney's Office. There were 107 individuals in attendance, representing 67 organizations and institutions such as Tanium, Indiana Department of Homeland Security, Rose-Hulman Institute of Technology, Eli Lilly and Company, Barnes & Thornburg, Purdue University, and the IN-ISAC Security Operations Center. Additional information about the Summit can be found at: https://cacr.iu.edu/events/cybersecurity-summit/index.php

CACR Helps Educate the Next Generation of Cybersecurity Professionals.

https://cacr.iu.edu/news/2016/CACR-Security-Matters-Cybercamp.php

CACR continued its K-12 education and outreach effort by holding its second Security Matters Cybercamp for high school students this year. Attended by 9 students from Bloomington and Indianapolis and one from out of state (Illinois), this two-day event was designed from the ground up to expose youth to professional cybersecurity concepts in a format that interested and challenged them. This year's camp stressed the importance of online security and privacy, provided insight into a hacker's mind, delved into topics such as ransomware and computer forensics, and gave students tools they can use to protect themselves against cybercrime. It also discussed potential cybersecurity careers and demonstrated physical cybersecurity through a tour of the IU Data Center.

CACR Helps Secure Protected Health Information.

CACR provided HIPAA consulting service to IU's University Information Technology Services through a custom, NIST-based risk management framework that secures protected health information end to end through a nuanced, workflow risk based approach. Through CACR's efforts, IU's has become one of the few central IT organizations nationwide that is compliant both with cybersecurity standards and with the HIPAA Security Rule and Centers of Medicare and Medicaid Services (CMS) data security requirements. This year, CACR helped University Information Technology Services add 8 new NIST/HIPAA compliant systems. This includes the Bomgar, Big Red II, Data Capacitor 2, Global ConfigMgr, Intelligent Infrastructure, OnBase, Slashtmp, and WebCAMP. CACR also engaged in over 80 HIPAA consultations and collaborative projects with researchers and IT professionals within the Indiana Clinical and Translational Sciences Institute and other IU areas; assisted Regenstrief Institute establish a NIST-based HIPAA security process, and trained over 350 University Information Technology Services staff on the HIPAA Security Rule.

On the national front, CACR provided 6 HIPAA consultations to other academic and other organizations and presented its NIST-based framework at the American Medical College Conference on Privacy and Security.

Educating Journalists

https://cacr.iu.edu/events/2017/cybersec-for-jounalists.php

Journalists are increasing need private communications to protect their sources and work in progress. Along with colleagues from the Electronic Frontier Foundation, the IU Center for International Media Law and Policy Studies, the Indianapolis Star, and the IU Maurer School of Law, CACR hosted a panel to explore and educate journalists on cybersecurity to support their needs.

Leadership & Staff

Leadership·

- CACR Director Von Welch has more than a decade of experience developing, deploying, and providing cybersecurity for private and public sector HPC and distributed computing systems.
- Administrative Director, Leslee Bohland has over two decades of accounting and financial management experience.
- Chief Policy Analyst Craig Jackson is a co-PI on the NSF Cybersecurity Center of Excellence and lead for CACR's collaboration with Naval Surface Warfare Center Crane Division.
- Associate Director Scott Orr, School of Science at IUPUI, leads the coordination of Indiana University's designation as a Center for Academic Excellence in Information Assurance/Cyber Defense Education and Center for Academic Excellence Information Assurance/Cyber Defense Research.
- Senior Systems Analyst Susan Sons leads CACR's work with Open Science Grid cybersecurity and the Internet Civil Engineering Institute.

Staff: CACR staff help manage the daily operations of the Center. CACR staff includes administrative, management, external relations support, as well as security and policy analysts. Current staff includes:

Diana Borecky, Senior Administrative Assistant
Ryan Kiser, IT Specialist
Scott Russell, Senior Policy Analyst
Mark Krenz, Lead Security Analyst
Randy Heiland, Senior Systems
Zalak Shah, Systems Analyst
Anurag Shankar, Senior Security Analyst
Amy Starzynski Coddens, Education, Outreach and Training Manager

Fellows and Key Liaisons: CACR has a dozen fellows, each one bringing unique insight to the Center and connecting to the center, allowing it to capitalize on the interdisciplinary strengths of Indiana University and the broader community. Fellows represent a wide range of perspectives, including law, policy, ethics, and informatics. Current fellows are as follows:

Fred H. Cate, Maurer School of Law
L. Jean Camp, School of Informatics and Computing
Jake Chen, School of Informatics and Computing (IUPUI)

Robert Cowles, Brightlite Information Security

Rachel Dorkery, Maurer School of Law

Arjan Durressi, Department of Computer and Information Science (IUPUI)

David P. Fidler, Maurer School of Law

Apu Kapadia, School of Informatics and Computing

Steven Myers, School of Informatics and Computing

Scott J. Shackelford, Kelley School of Business

Robert Templeman, Naval Surface Warfare Center, Crane Division

Joseph Tomain, Maurer School of Law

Xiaofeng Wang, School of Informatics and Computing

Xukai Zou, Department of Computer Science (IUPUI)

Associate Director William K. Barnett is the Indiana CTSI and Regenstrief Chief Research Informatics Officer. Associate Director Mark Bruhn is Indiana University's Associate Vice President for Assurance and Public Safety.

For More Information About CACR

The latest information about CACR, please visit https://cacr.iu.edu To discuss collaboration with CACR on addressing your challenges, contact cacr@iu.edu .

Acknowledgments

CACR's work is funded by the IU Office of the Vice President for Information Technical, the IU Office of the President, the Department of Homeland Security, and the National Science Foundation (grants 1547272, 1642070, 5107311). None of the opinions expressed in this report should be assumed to represent the opinions of funding entities.

Science Gateways Research Center

The Science Gateways Research Center at IU researches, develops, and operates science gateways in collaboration with many clients and partners. SGRC was created in 2016 as part of the National Science Foundation funded Science Gateways Community Institute (NSF grant #1547611). The institute is lead by San Diego Supercomputer Center and partners with researchers at six different institutions, including IU. The purpose of the partnership and the Science Gateways Research Center, in particular, is to accelerate the development and application of sustainable science gateways that address the needs of a broad spectrum of researchers.

Incubator PY1 highlights

Incubator

- Bootcamp 1
  - April 24–28, 2017
  - Purdue Research Park, Indianapolis
  - 23 attendees
- Bootcamp 2
  - October 2-6, 2017
  - Purdue Research Park, Indianapolis
  - Accepting up to 10 projects to participate as a cohort
  - http://sciencegateways.org/upcoming-events/science-gateways-bootcamp/
- Consulting Engagement Requests
  - 3 active engagements, more in pipeline
  - http://sciencegateways.org/request-services

Bootcamp at a Glance

- Full 5 days
- Knowledge dissemination
- Interactivity
- Community formation
- Putting away the normal daily routine
- Homework

8-Week Internships

Summer internships offered each year for students interested in developing their gateway skills. Participants are placed at one of the SGCI or partner/client sites. **6 students placed during summer 2017.**


Hagen Hodgkins and Joel Gonzalez-Santiago
Purdue University


Thomas Johnson III and Disaiah Bennett
Indiana University

Jacob Harless
College of William & Mary

Tatyana Matthews
Texas Advanced Computing Center (TACC)

4- Week Coding Institute at ECSU

- 8 Undergraduates and 2 graduate students
- Trainers provided by Indiana University and the Software Carpentry Institute
- Python, UNIX Shell, Git, R-programming

4- Week Coding Institute at ECSU

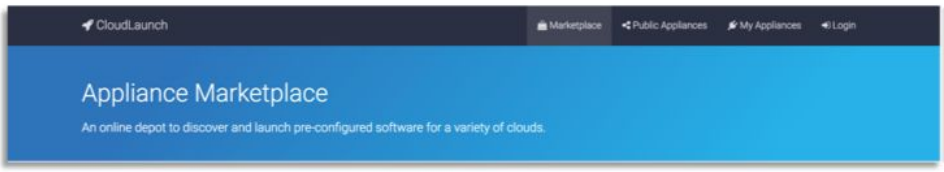
- 8 Undergraduates and 2 graduate students
- Trainers provided by Indiana University and the Software Carpentry Institute
- Python, UNIX Shell, Git, R-programming

University of South Dakota Campus Gateway

PI: Doug Jennewein, University of South Dakota

Consultant: Eroma Abeysinghe, Indiana University

Galaxy Cloud Launch



PI: Enis Afgan, Johns Hopkins University

Consultant: Marcus Christie, Indiana University

Enabling Science & Broader Impacts

SGCI

Digital Science Center

Research in Digital Science Center

Geoffrey Fox, August 25, 2017
Digital Science Center
Department of Intelligent Systems Engineering
gcf@indiana.edu, http://www.dsc.soic.indiana.edu/,
http://spidal.org/

- Judy Qiu, David Crandall, Gregor von Laszewski, Dennis Gannon
- Supun Kamburugamuve, Pulasthi Wickramasinghe, Hyungro Lee, Jerome Mitchell
- Bo Peng, Langshi Chen, Kannan Govindarajan, Fugang Wang
- Internal collaboration. Biology, Physics, SICE
- Outside Collaborators in funded projects: Arizona, Kansas, NIST, Purdue, Rutgers, San Diego Supercomputer Center, Stanford, SUNY Stony Brook, Virginia Tech, Univ. of Tennessee Knoxsville, UIUC, Utah

Digital Science Center

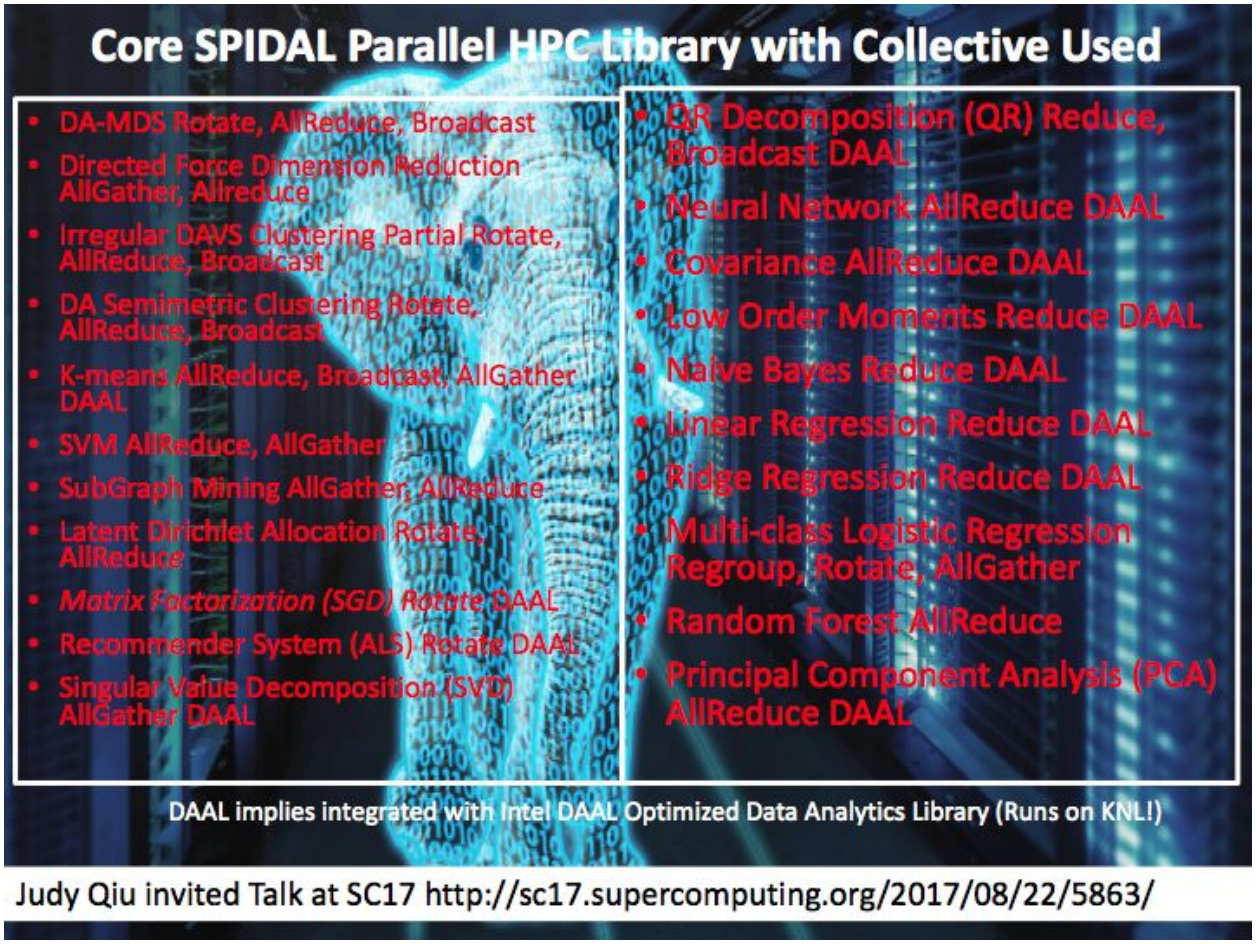
- Big Data Engineering (Data Science) and parallel computing research with technology and applied collaborators
  - System architecture, performance
- Run computer infrastructure for Cloud and HPC research
  - 64 node system **Tango** with high performance disks (SSD, NVRam = 5x SSD and 25xHDD) and Intel KNL (Knights Landing) manycore (68-72) chips. Omnipath interconnect
  - 128 node system **Juliet** with two 12-18 core Haswell chips, SSD and conventional HDD disks. Infiniband Interconnect
  - 16 GPU, 4 Haswell node deep learning system **Romeo**
  - All can run HDFS and store data on nodes
  - 200 older nodes for Docker, OpenStack and general use
- Teach basic and advanced Cloud Computing and bigdata courses
- Supported by Gary Miksik, Allan Streib, Laura Pettit (partial)

DSC Research Activities

- Building SPIDAL Scalable HPC machine Learning Library
- Applying current SPIDAL in Biology, Network Science, Pathology
- Polar (Radar) Image Processing (Crandall)
- Data analysis of experimental physics scattering results
- Work with NIST on Big Data Standards and non-proprietary Frameworks
- Integration of Clouds&HPC; Big Data&Simulations (international community)
- Harp HPC Machine Learning Framework (Qiu)
- Twister2 HPC Event Driven Distributed Programming model
- IoTCloud. Cloud control of robots – licensed to C2RO (Montreal)
- Cloud Research and DevOps for Software Defined Systems (von Laszewski)
- Network for Computational Nanotechnology – Engineered nanoBIO Node

Engineered nanoBIO Node NSF EEC-1720625

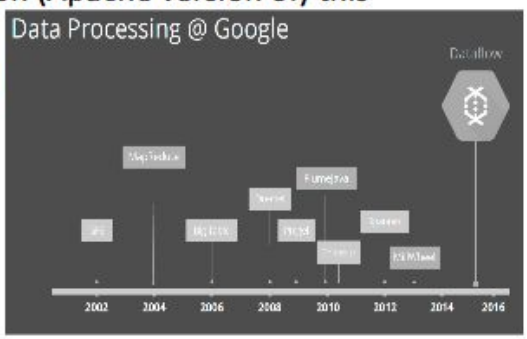
- Starts September 1, 2017 $4M over 5 years
- Involves ISE Bioengineering (Macklin, Glazier) and Nanoengineering (Jadhao) faculty, Chemistry (Douglas to help code validation), PTI Gateway group, Purdue and UIUC as separately funded nanoMFG node
- Deploys simulation and data analysis (from SPIDAL) tools on nanoHUB with a strong emphasis on backend parallel codes with multiscale physics
  - Functional nanoparticles and nanostructures with user-selected properties,
  - Nanodevice-cell interactions and phenotype links,
  - Enable the engineering of multicellular systems
  - Visualize trajectories in simulations
- Strong Education and Research components and will fund outreach and software engineering



Judy Qiu invited Talk at SC17 http://sc17.supercomputing.org/2017/08/22/5863/

Components of Big Data Stack

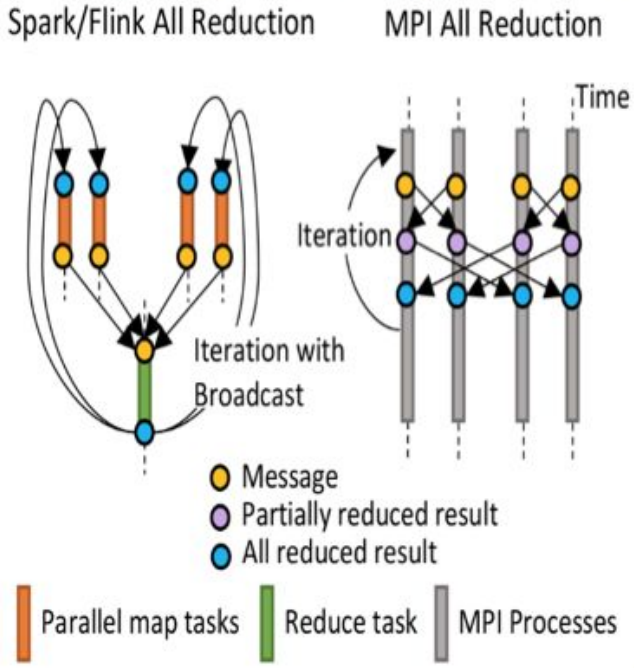
- Google likes to show a timeline; we can build on (Apache version of) this
- 2002 **Google File System** GFS ~HDFS (Level 8)
- 2004 **MapReduce** Apache Hadoop (Level 14A)
- 2006 **Big Table** Apache Hbase (Level 11B)
- 2008 **Dremel** Apache Drill (Level 15A)
- 2009 **Pregel** Apache Giraph (Level 14A)
- 2010 **FlumeJava** Apache Crunch (Level 17)
- 2010 **Colossus** better GFS (Level 18)



- 2012 **Spanner** horizontally scalable NewSQL database ~CockroachDB (Level 11C)
- 2013 **F1** horizontally scalable SQL database (Level 11C)
- 2013 **MillWheel** ~Apache Storm, Twitter Heron (Google not first!) (Level 14B)
- 2015 **Cloud Dataflow** Apache Beam with Spark or Flink (dataflow) engine (Level 17)
- Functionalities not identified: **Security(3), Data Transfer(10), Scheduling(9), DevOps(6), serverless computing** (where Apache has **OpenWhisk) (5)**

HPC-ABDS Levels in ()

HPC Runtime versus ABDS distributed Computing Model on Data Analytics



Hadoop writes to disk and is **slowest**; **Spark** and **Flink** spawn many processes and do not support AllReduce directly; **MPI** does in-place combined reduce/broadcast and is **fastest**

Need Polymorphic Reduction capability choosing best implementation

Use HPC architecture with
Mutable model
Immutable data

Some Important Trends

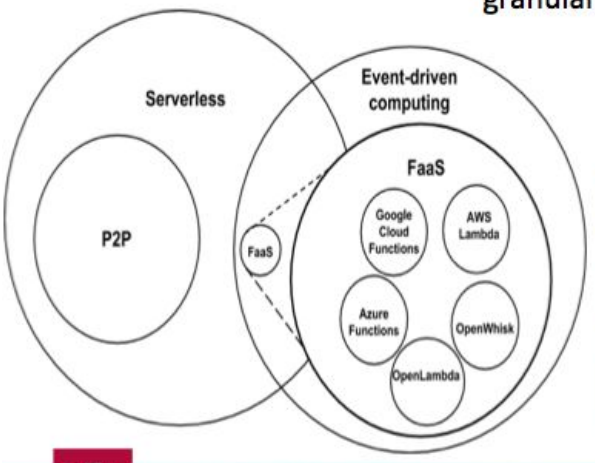
- Rich software stacks:
  - **HPC** (High Performance Computing) for Parallel Computing
  - **Apache** for Big Data Software Stack ABDS including some edge computing (streaming data)
- **Parallel and distributed computing** have different requirements but often similar functionalities
  - Apache stack ABDS typically uses distributed computing concepts
  - For example, Reduce operation is different in MPI (Harp) and Spark
- Big Data requirements are not clear but there are a few key use types
  1) Pleasingly parallel processing (including **local machine learning LML**)
  2) **Database model** with queries again supported by MapReduce for horizontal scaling
  3) **Global Machine Learning GML** with single job using multiple nodes as classic parallel computing
  4) **Deep Learning** certainly needs HPC – possibly only multiple small systems
- Current workloads stress 1) and 2) and are suited to current clouds and to ABDS (with no HPC)
  - This explains why Spark with poor GML performance is so successful
- **Serverless (server hidden) computing attractive to user**:
  "No server is easier to manage than no server"

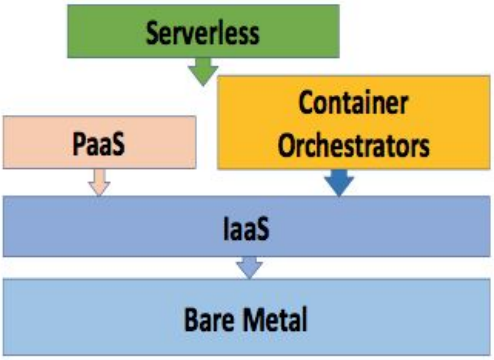
Hopefully will change

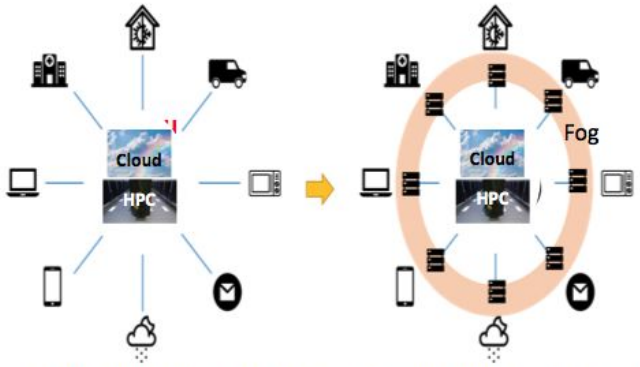
Event-Driven and Serverless Computing

Note GridSolve was FaaS

- Cloud-owner Provided Cloud-native platform for
- Short-running, Stateless computation and
- Event-driven applications which
- Scale up and down instantly and automatically and
- Charges for actual usage at a millisecond granularity







Centralized HPC Cloud + IoT Devices

Centralized HPC Cloud + Edge = Fog + IoT Devices

HPC Cloud can be federated

Implementing Twister2 to support a Grid linked to an HPC Cloud

Twister2: "Next Generation Grid - Edge – HPC Cloud"

- Original 2010 **Twister** paper has 878 citations; it was a particular approach to MapCollective iterative processing for machine learning
- **Re-engineer** current Apache Big Data and HPC software systems as **a toolkit**
- Support a **serverless (cloud-native) dataflow event-driven HPC-FaaS (microservice)** framework running across application and geographic domains.
  - Support all types of Data analysis from GML to Edge computing
- Build on Cloud best practice but use HPC wherever possible to get high performance
- Smoothly support current paradigms Hadoop, Spark, Flink, Heron, MPI, DARMA …
- Use **interoperable** common abstractions but multiple **polymorphic** implementations.
  - i.e. do not require a single runtime
- Focus on Runtime but this implies HPC-FaaS programming and execution model
- This defines **a next generation Grid** based on data and edge devices – not computing as in old Grid

See long paper http://dsc.soic.indiana.edu/publications/Twister2.pdf

Data to Insight Center

The Data To Insight Center (D2I) engages in interdisciplinary research and education in the preservation of scientific data, big data, cyberinfrastructure, frameworks for distant reading in digital humanities, provenance, and cloud computing. It carries out data modeling for large and complex data, pioneers new forms of data publishing including tools for embedded data curation early in the lifecycle of data and models for persistent identifier user; innovates through new tools for provenance capture; pioneers novel big data cyberinfrastructure; studies communities of practice in data sharing, and supports early career researchers through a fellows program with the Research Data Alliance.

In 2017, D2I continued to advance its major large-scale projects, such as the HathiTrust Research Center, and engaged in several new projects that are already marked with several milestones. The detailed descriptions of D2I 2017 achievements are below.

The HathiTrust Research Center (HTRC)

During this year, the HTRC Data Capsule service - a service that enables secure analytical access to HathiTrust Digital Library via a virtualized environment - has undergone considerable enhancements. HTRC infrastructure can now support a larger number of simultaneous users, and each individual user can use a much more powerful environment for their analysis. New tools developed to work within Data Capsule allow users to quickly import full-text volumes and metadata, start natural language processing, explore and visualize topics, and develop Python tools via Python software development kit (SDK).

In addition, HathiTrust provided strong support of HTRC with significant investment in hardware equipment costing approximately $102,000. This new hardware will allow use of in-copyright content that will be available in HTRC 4.0 release. Productive collaboration with UITS Research Technologies High Performance Systems group in hardware acquisition and system administration has made the enhancements smooth and efficient. HTRC 4.0 will be hosted on the new dedicated cluster managed by HPS, with the enhanced security infrastructure and user interfaces. The phased release is expected to begin Fall 2017, providing the community with non-consumptive access to full HathiTrust corpus.

In Spring 2017 D2I has been awarded a National Leadership Grant for Libraries from the Institute for Museum and Library Services. The grant titled “Data Capsule Appliance for Research Analysis of Restricted and Sensitive Data in Academic Libraries” ($320,546, #LG-71-17-0094-17) will enhance HTRC Data Capsule, enable partner libraries to use it in providing access to other, non-HathiTrust collections, and examine how such infrastructure can be developed in a collaborative manner through the framework of participatory design.

HTRC has funded 4 different Advanced Collaborative Support (ACS) Proposals during this reporting timeframe:

- **"Fighting Fever in the Caribbean: Medicine and Empire, 1650-1902,"** Mariola Espinosa, University of Iowa: This project seeks to explore the history of yellow fever in the Caribbean by comparing how the disease was described by residents of the Caribbean to the European perspective, including through sentiment analysis of text referencing yellow fever. Her work will be visualized spatially in a map generated with support from the University of Iowa's Digital Scholarship and Publishing Studio. She will build a corpus of text from the HathiTrust Digital Library related to yellow fever and filth in the Caribbean to track the use of the terms "filth" and "filthiness" from 1650 to 1902.

- **"Inside the Creativity Boom,"** Samuel Franklin, Brown University: This project will map the increasing use and shifting meanings of the words "creative" and "creativity," with a particular focus on the twentieth century. A custom "creativity corpus" will be assembled and processed to identify linguistic patterns via a number of text analysis and natural language processing techniques. Brown's project will make use of the functionality developed for HathiTrust + Bookworm.

- **"The Chicago School: Wikification as the First Step in Text Mining in Architectural History,"**Dan Baciu, Illinois Institute of Technology: This project will look at the Chicago School of architecture and examine its history of reception over the last 75 years, as well as identify patterns in its international spread and influence. Baciu will use named entity recognition in his analysis, notably deploying the Wikifier tool on a large sample corpus of HathiTrust data for the first time.

- **"Signal and Noise and Pride and Prejudice: Toward an Information History of Romantic Fiction,"** Dallas Liddle, Augsburg College: This project will test two hypotheses about information theory and information density as they relate to a digital humanities approach to studying Romantic-era British fiction. The concept of "information" used in mathematical information theory may help digital humanists evaluate the information density of textual forms. This project tests a theory that the popular and critical success of the novel in British print culture after 1815 may be related to increased information density and sophistication of information encoding in those years, especially via innovations introduced by Jane Austen and Walter Scott.

HTRC will host our 4th UnCamp at the University of California Berkeley in January 2018.

SEAD

The SEAD project is coming to end with the final update to be deployed in Summer 2017. SEAD publishing suite is now a fully operational set of services that can support individuals and small teams in their curation and publishing needs. The software built under SEAD will transition to operate under the National Data Service (NDS), thereby contributing to the emerging national data infrastructure.

Major components of SEAD publishing pipeline are openly available for download and use via open source licenses. In particular, the D2I team has released an improved and new version of the Matchmaker, a framework for matching Research Objects, Repositories and People during the data curation and publication process. A video and examples accompanied the software release.

Research Data Alliance (RDA)

D2I is involved in several initiatives with the Research Data Alliance (RDA, rd-alliance.org), including building RDA presence in the US through leadership and engagement and advancing early career researchers and professionals in RDA through the Data Share fellowship program. This year D2I has been also successful in bringing RDA outputs into broader adoption use.

One of the successes is the adoption of RDA outputs by organizations - members of the Pacific Rim Application and Grid Middleware Assembly (PRAGMA), where D2I plays an active role. Specifically, the D2I team leveraged two recent recommendations from RDA - Persistent Identifier Information Type (PIT) and Data Type Registry (DTR), to bring persistence to the results of genomic analyses performed by the International Rice Research Institute (irri.org). Persistent IDs and registration of data objects will enhance sharing of those objects. The services developed in this adoption project are designed to be reusable in other cases.

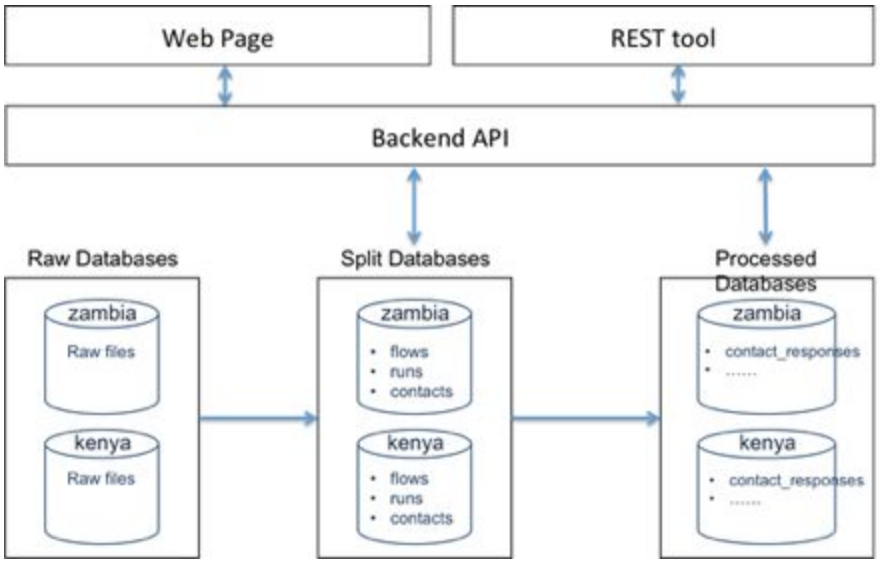
The second success is a grant received by D2I from the National Science Foundation titled “Robust Persistent Identification of Data (RPID)” ($199,047, #1659310). In collaboration with IU Research Technologies and two other institutions D2I will establish and support a testbed for minting and solving persistent identifiers (PIDs) to stimulate and enable evaluation of new outputs of the Research Data Alliance (RDA) in PID-oriented data management. The testbed will be responsive to priorities in science and education, specifically as part of the cyberinfrastructure ecosystem for data-intensive research.

Smallholder Farmer Data

In 2017 D2I actively collaborated with the team of researchers from the IU department of geography to improve and automate the practices of survey data collection and preservation. We developed an approach to long-term management and preservation of mobile SMS survey data as part of a larger project that

examines adaptation of small-scale farmers in Africa to climate change. An automated pipeline ingests weekly data from a cloud-based platform to local data servers, while maintaining security and confidentiality (see Fig. 1 below).

The pipeline provides an initial infrastructure that can be further developed to curate, monitor, and visualize data health and trends.



Midwest Big Data Hub (MBDH)

MBDH is led out of NCSA with co-leads at IU, University of Michigan, Iowa State, and University of North Dakota. Co-PI Plale leads the Data Science Ring; Co-PI Bernice Pescosolido leads the Network Science Spoke. Through MBDH D2I received Microsoft Azure credits to test data publishing and persistent identification in the cloud for hands-on workforce training in big data methodologies. The project called SEADTrain uses SEAD publishing tools to publish data into the Azure Cloud and implements RDA recommendations in persistent IDs to assign minimal metadata and quickly sort through complex heterogeneous objects, such as streaming data from environmental sensors. The first training session using SEADTrain will be held in July 2017 at the ESIP meeting in Bloomington, IN.

National Center for Genome Analysis Support

What are the major goals of the project?

The major goals of the NSF ABI Sustaining Award are to support National Center for Genome Analysis' (NCGAS) continuing and expanding activities during this award's duration, including:

1. Provide excellent bioinformatics consulting services, to all NSF-funded researchers in need.
2. Maintain, support, and deliver genome assembly and analysis software on national CI systems.
3. Provide education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data
4. Disseminate tools for genome assembly and analysis.
5. Provide long-term archival storage for genome biologists.

Emphasis is placed on genome and transcriptome assembly at the technically challenging end of the spectrum of current bioinformatics—for example *de novo* genome and transcriptome assembly—where both specialized computational resources and applications are needed.

Major Activities

NCGAS is a collaborative project between the lead institution, Indiana University (IU), and the Pittsburgh Supercomputing Center (PSC) at Carnegie Mellon University. At IU, NCGAS is part of the Indiana University Pervasive Technology Institute and has significant HPC facilities, human resources, and administrative support from IUPTI and IU. Likewise, NCGAS-funded collaborator PSC maintains extensive HPC resources and supporting services. During the second year of this sustaining award, NCGAS has continued to make significant strides in using NSF funding—with additional funding and facilities from IU and PSC—to aid discovery and innovation in the biological sciences in the US. Under the direction of IU, NCGAS has developed new opportunities through collaborative efforts between IU and PSC and has continued to aid in discoveries that range from a better understanding of basic biological processes, to discoveries that will aid management of economically and ecologically important animals and plants.

How national is the National Center for Genome Analysis Support? We now server researchers in 41 states (Fig. 1a), including 12 EPSCoR states, and in partnership with the Trinity development team (see Synergistic activities) have users around the world (Fig. 5).

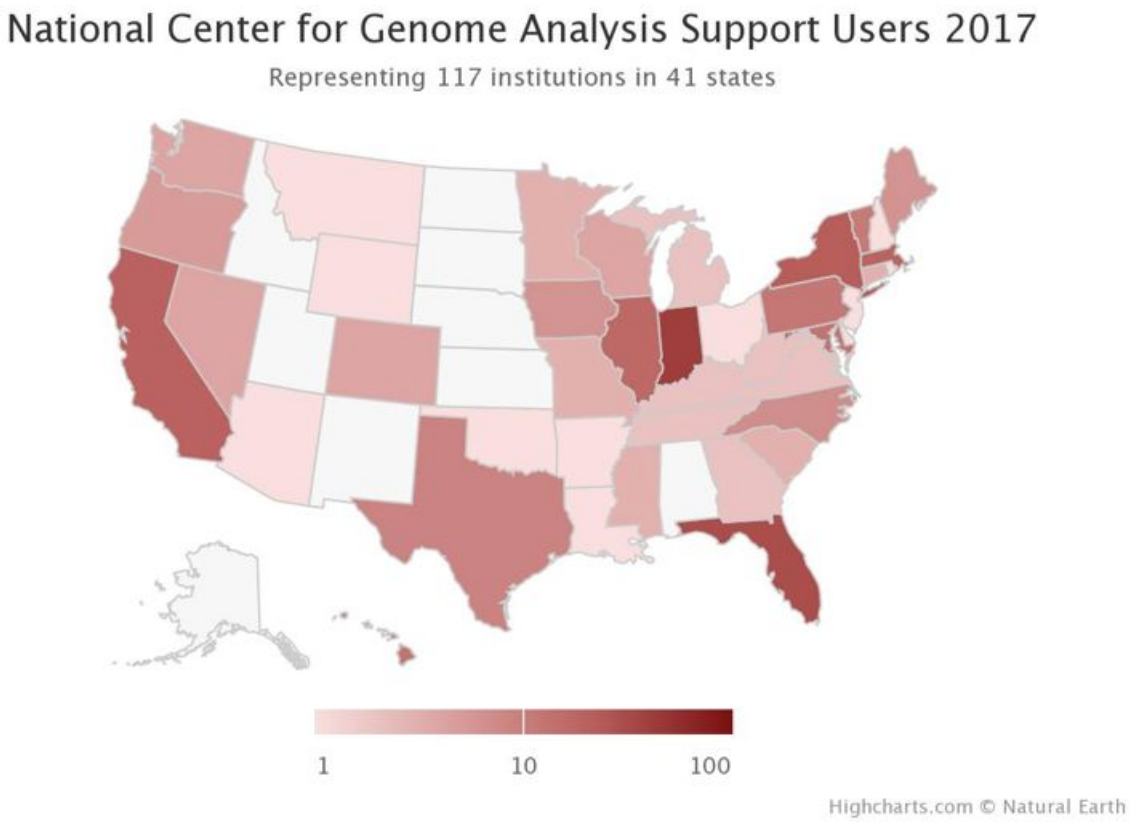


Figure 1. States with NCGAS Users

Provide excellent bioinformatics consulting services.

NCGAS' most significant accomplishments in support of biological and bioinformatics research continue to be in researchers' discoveries from RNA and DNA transcriptome, metatranscriptome, genome, and metagenome assemblies.

In year 2, NCGAS aided researchers in completing many *de novo* assemblies and genomic analysis, including the following organisms (completed and on-going):

- Coffee, peanut and sweet potato transcriptomes
- Daphnia genomes (population genomics and *de novo* assemblies)
- Barred tiger salamander transcriptomes
- Diatoms transcriptomes
- The diverse microbial clade Stramenopila + Alveolata + Rhizaria (SAR) (*de novo* genomes, transcriptomes)
- Heliconius butterflies transcriptomes
- Carrion Flies (forensic arthropods), *de novo* genome assembly/resequencing, transcriptomes
- Mussels,
- Crawfish Frog (endangered)
- Bahama Giant conch (a mollusk)
- Little Brown Skate,
- D. galeata

In addition, NCGAS supported 223 biologists doing research in the general area of genome analysis (15 named new allocations) during the current year's funding. Assistance has been provided through 386 short consultations and 29 extended consultations. Details regarding many of the extended consultations are provided in an attachment.

Maintain, support, and deliver genome assembly and analysis software on national CI systems

NCGAS continues to assist NSF researchers in genomics research. From the beginning we have accomplished this by forming a "supply line" from the researchers' specific data and questions to HPC hardware, specialized applications and knowledge. Some researchers only need access to large memory clusters, which we can provide in a number of ways; others need instruction in basic HPC use and genomic analysis. We have been successful in this and continue to attract new users, often by word-of-mouth (documented in our just closed survey of users). One of our on-going tasks is to stay current: new hardware becomes available, state-of-the-art applications change, new data types become available (we are starting to see a significant number of PacBio data sets), and researchers change. For example, this year we installed or updated five packages (spades, hisat2, salmon, STAR, kallisto) for assembly and analysis of PacBio long-read sequencing data. Hybrid assemblies are quickly becoming popular, and we installed Canu, MaSuRCA and PacBio's SMRT analysis software. Funded collaborator PSC also makes many of these packages available on PSC systems, and also focuses on enabling high-quality metagenome assembly and analysis (see PSC report).

NCGAS at IU provides accounts to multiple clusters for direct command-line access:

- The large memory *Mason* IU cluster
- The new *Carbonate* IU cluster (will replace Mason shortly)
- The *Jetstream* cloud environment
- Additional XSEDE resources, including PSC's *Bridges*, through an NCGAS XSEDE Community Allocation

NCGAS at IU also provides access to bioinformatics software through online web (graphic) portals:

- NCGAS Galaxy web portal: providing access to the widely used Galaxy workflow system on Mason and other XSEDE-supported resources
- Trinity RNA-Seq Galaxy portal: running on IU's Karst cluster
- GenePattern Analysis Package, running on IU's Karst

We now support public and private genome browsers for: XXX, XXX, etc. Overall, we have installed or updated 44 software packages across the systems described above (see attachment describing significant software activities).

Disseminate tools for genome assembly and analysis

NCGAS has supported the creation of the "XSEDE National Integration Toolkit" (XNIT). XNIT is a suite of software available for download and installation on computational clusters. NCGAS has added whole

suits of bioinformatics software supported by NCGAS on XSEDE in the past, but this activity has lapsed in the last year due to personnel shortage and more pressing priorities. We hope to return to this important activity in the second year.

Provide long-term archival storage for genome biologists
NCGAS and IU continue to provide access for all NCGAS users to IUScholarWorks, a digital repository provided by the IU Libraries for showcasing and preserving research findings, and the Scholarly Data Archive (SDA), which provides extensive capacity (approximately 42 PB of tape) for storing and accessing research data.

Provide education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data.
While the bulk of NCGAS personnel's time is devoted to one-on-one consultation and assistance, we also do outreach and trainings (see 1.3 and 5.2). For example:

- Sheri Sanders has just finished teaching at the MDI Biological Laboratory's 2016 Environmental Genomics course and we are starting to plan for presentations at the 2017 PAG meeting.
- We started a collaboration with Raphael Isokpehi at Bethune-Cookman University, providing student access to Mason and teaching classes remotely. This relationship will expand in the future.

The most important training and professional development activities have been presentations and tutorials provided at national and international conferences. A summary of these tutorials is provided below:

- 200 participants in tutorials (~500 contacts at events). In the last year, some of the events attended were:
  - Galaxy Community Conference 2016/Bloomington IN (we hosted);
  - GMOD User Community 2016 / Bloomington IN (we hosted);
  - Extreme Science and Engineering Discovery Environment (XSEDE) 2016 / Miami FL
  - Plant and Animal Genomes 2017. San Diego.
  - MDI Biological Laboratory's 2016 Environmental Genomics course / Bangor ME
  - Cyber-lectures to bioinformatics students at Bethune-Cookman.

Specific Objectives

The software supported by NCGAS as of the end of the first four years of NSF funding includes 44 packages described in detail in the attached file on significant software activities.

Significant results

Provide online help, consulting, and tutorials related to genome analysis.

Key highlights of NCGAS support include:

- Consulting. NCGAS in the current year completed a total of 386 short term consulting engagements (those taking less than 4 hours of staff time to resolve) and 29 long term consulting engagements (taking more than 4 hours of staff time to resolve). Many long term consultations are research collaborations that last months or years, with NCGAS staff becoming partners, playing a critical role in discoveries by scientists receiving NCGAS help.
- NCGAS completed tutorials and training and outreach activities attended by hundreds of attendees (see 1.3).
- The past year has included the Jetstream cloud opening. NCGAS has worked to bring biologists to this NSF resource, including: 1) helped researchers at the University of Arkansas Fayetteville to establish, provision, and use Jetstream VMs, to complete analysis of both northwest endangered river fish species, and the distribution of rattlesnake species; 2) Established genome browsers to, that link to Wrangler for storage; 3) Adding the research group of Tim Grimes to establish proteomic galaxy instances for both research and teaching purposes.
    - In another capacity, NCGAS is working with groups funded by Information Technologies in Cancer Research (ITCR) to use Jetstream in their work.

Consultant services are provided by telephone, email (a ticketing system tracks requests), and in-person consultations. Consulting hours are typically 8 am to 5 pm weekdays, but support activities often extend beyond local business hours when there is time pressure on a researcher.
In the last year NCGAS engaged in a total of 15 new long-term projects, described in the attached file on consulting and significant results.

A significant activity is the strengthening of our partnership with PSC (see 1.5). With Philip Blood and PSC as a funded member of the NCGAS team in the current grant, we have been able to escalate our relationship. Activities include: 1) coordination of software suites; 2) increasing use of Bridges when very large memory nodes are needed; 3) establishing a shared Luster file system, allowing increased interoperability (under way). 4) beginning to build a center for metagenomic analysis centered at PSC. While offering a full metagenomics service is beyond the capacity of the current grant, we are establishing foundations. We plan to submit an ABI Development grant this fall to further this work: Blood would server as PI, with IU playing a secondary role, but it will be an NCGAS-branded service. The division of labor between the two center is still to be established.

Key outcomes or other achievements

The key outcome during the second year of this sustaining award is the continued success of NCGAS in delivering an effective consulting service focused on accelerating the research of biologists and bioinformaticians, and in so doing accelerated biological discoveries in the US. NCGAS provides a robust "supply chain" from NSF-funded and other supercomputers, through specialist applications and knowledge, to bench and field scientists across the country. NCGAS' ongoing efforts have helped enable 8 peer-reviewed scientific publications that have been published in 2016-2017 (beyond those reported in the first year's report).

Results of the 2017 user survey

We have just completed a user satisfaction survey, with 56 users responding. Much like our last survey (about 2 years ago), we found there was broad satisfaction with our services, and that NCGAS had been essential to many users' research (Fig. 2). If there were complaints, they seemed to be focused on areas where we are limited by personnel—a limitation we are aware of. In the last survey, we didn't ask to correct question, to understand what funding sources our researchers relied on. We find that NCGAS is now used primarily by NSF-funded researchers (Fig. 3. After further analysis, we will generate and make available a white paper reporting the results of this survey.

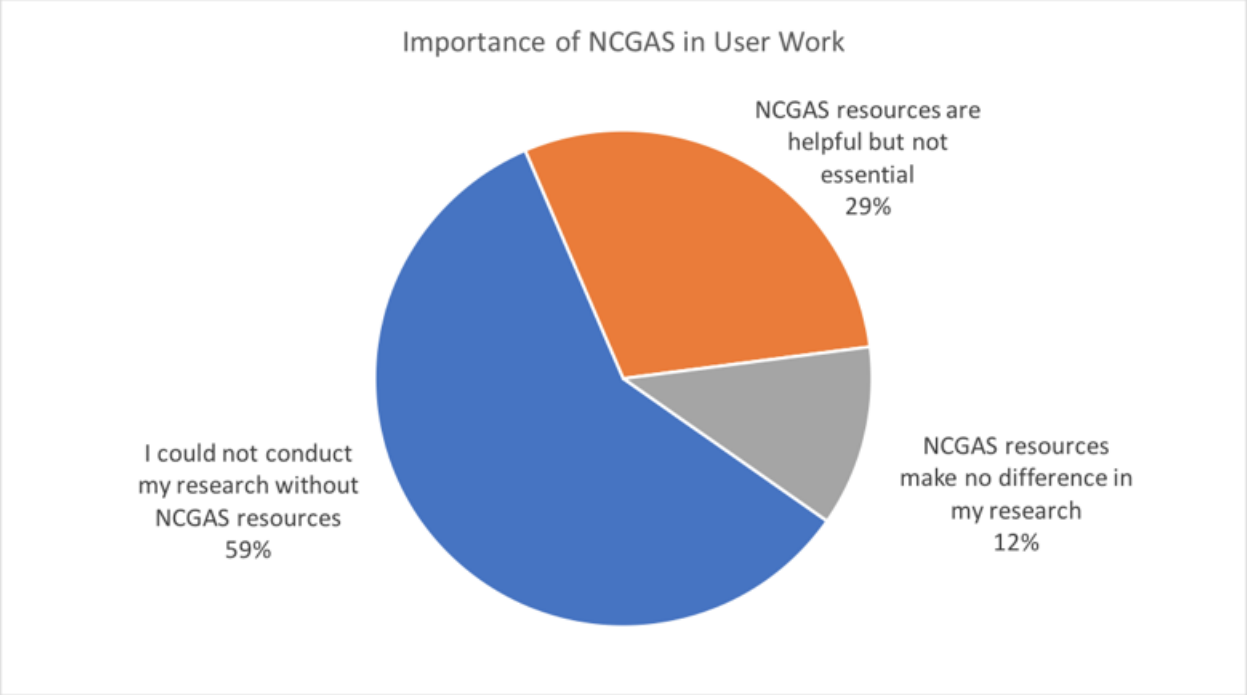


Figure 2. Usefulness

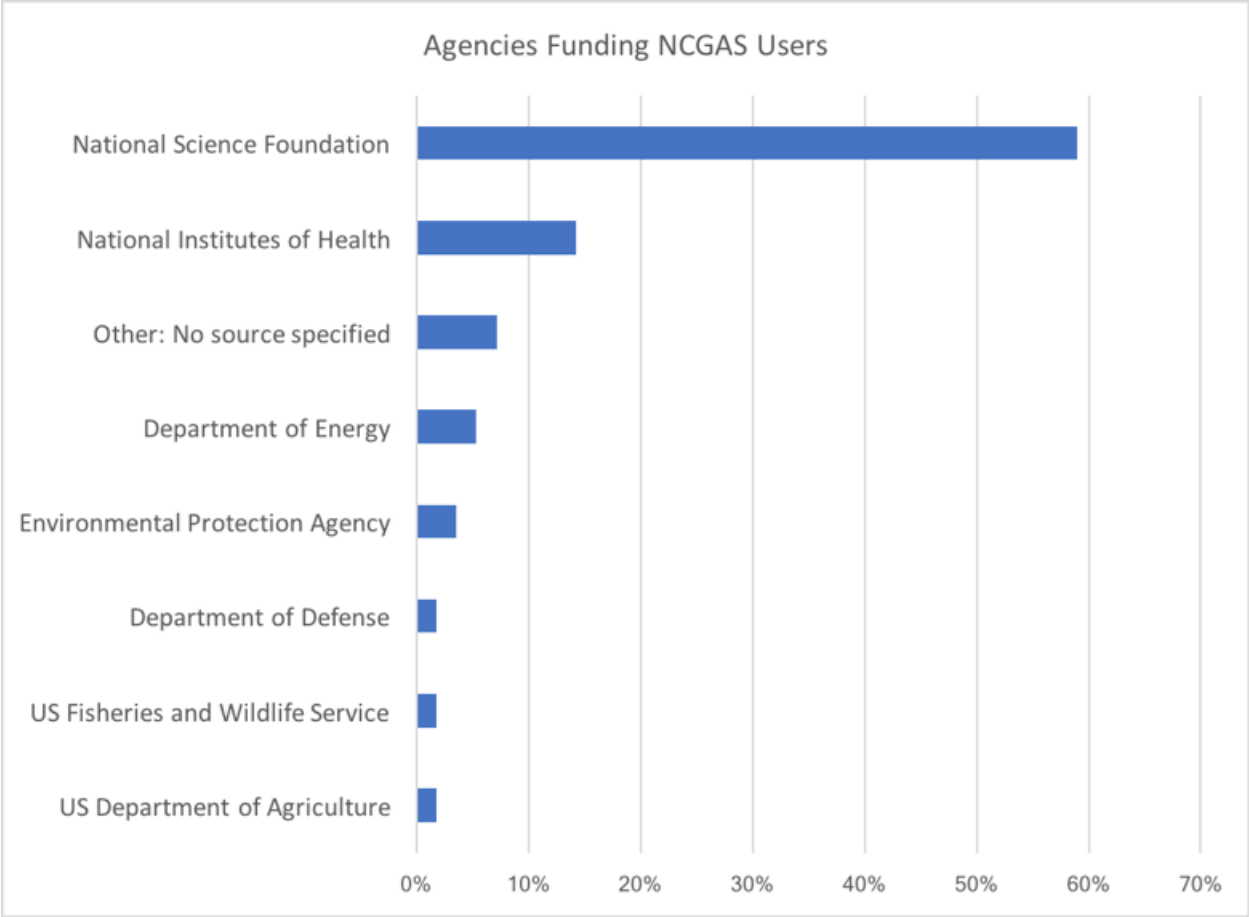


Figure 3. Funding sources reported by NCGAS to users

What opportunities for training and professional development has the project provided?

Dr. Thomas Doak was, during the first Development award, a postdoctoral fellow. Dr. Doak was promoted to the rank of Assistant Scientist at Indiana University, and is now PI and manager of this NSF sustaining award to continue and grow NCGAS services. Dr. Doak will be the primary author on next year's sustaining renewal.

Staff member Carrie Ganote is continuing her PhD program in bioinformatics, while in the employ of NCGAS. We have promoted her as a player in the international Galaxy community and she was on the organizing committee for the 2016 Galaxy conference, hosted at IU, and is now on the Galaxy Financial Committee. Ms. Ganote oversees many projects, and mentored Sheri Sanders. She has just been promoted from IT3 to IT4, reflecting her essential role in our organization.
Staff member Sheri Sanders has now been a NCGAS team member for a year, since finishing her PhD at Notre Dame, where she used transcriptomics to characterize salamander species, some endangered. Dr. Sanders' goal in joining NCGAS was to grow her understanding of IT and HPC, and how they impacted

the biological community. She now leads our efforts to support genome browsers and play a role in the GMOD development community.

Staff member Bhavya Nalagampalli Papudeshi will start as an NCGAS employee June 15th, 2017. Ms. Nalagampalli Papudeshi has just completed her master's degree in bioinformatics at SDSU, working in the lab of Liz Dinsdale. Her work includes optimization of metagenome assembly and binning tools to reconstruct population genomes, and she will strengthen our metagenomics support and our collaboration with PSC. She is already working with a collaborator of ours, Rob Edwards at SDSU, on shared metagenomics projects.

How have the results been disseminated to communities of interest?

Results have been disseminated to communities of interest in a variety of ways, including:

- Publications in scientific journals
- Presentations
- Birds of a feather sessions at technical conferences
- Displays and booths at national and international technical conferences
- Articles in the lay press, most notably in Science Node, https://sciencenode.org
- NCGAS web site at ncgas.org
- NCGAS Twitter and Blog accounts
- In-person contacts
- Email list distribution
- Newsletters

What do you plan to do during the next reporting period to accomplish the goals?

Goals for the next year

We initiated several new directions in the sustaining grant's first year, which we have carried forward: 1) continue to deepen our expertise in—and offerings of—genome browsers, for searchers to organize and distribute their results: 2) aggressively pursue metagenomic projects/researchers in collaboration with Phil Blood at PSC. He has considerable experience in metagenomics research, and our newest employee Bhavya Nalagampalli Papudeshi also supports this effort; 3) continue our use of the new Jetstream environment to aid genomics researchers.

NCGAS infrastructure was inaugurated 6 yrs. ago with the IU purchase of the large-memory cluster Mason, each node having half a terabyte of RAM (Random Access Memory), specifically to support DNA genome assembly and as an XSEDE-allocated resource, again primarily for genomics research. Mason is now antiquated, and is being replacing. This will be accomplished in three ways: 1) Mason is being replaced with the Carbonate cluster. Carbonate is considerably faster, with a higher memory-to-core ratio than Mason; Carbonate is onsite, and will open to early users June 1st. 2) NCGAS will take advantage of the new NSF-funded PSC cluster Bridges (see PSC annual report). Bridges is already in use for metagenomic assemblies through PSC and NCGAS has an XSEDE allocation to enable our users to utilize Bridges at PSC. 3) The NSF-funded cloud environment Jetstream: while not providing very large

memory, NCGAS has already helping researchers accomplish genomics science on Jetstream (*ex*. ecological-genomics projects from AR), and we will continue to expand its uses. We have just started to use Wrangler to provide storage for Jetstream VMs.

Other goals for the year:

- Incorporate our newest hire into the team, and use this as further opportunity to expand our metagenomics expertise in collaboration with PSC.
- The IU/TACC Jetstream cloud environment opens up a range of possibilities, which we will continue to take advantage of.
- We are now actively providing genome browsers to users, starting with the GMOD-base G and JBrowsers, and will continue this.
- Having PSC and Phil Blood as a funded partner, we continue to develop our support for metagenomics/metatranscriptomics. PSC has had an emphasis in metagenomics and Blood has considerable experience working with researchers and developers. The first step is ongoing: assembling a comprehensive tool set and proving this to researchers. We will then move to a metagenomics Galaxy instance. We are particularly interested to see if we can serve a metagenomic component of the NEON project.

Pursue Field and Marine Stations as NCGAS and Jetstream clients

In conjunction with the Jetstream development team, we have just completed a survey of field and marine station directors and managers, as represented by the membership of the Organization of Biological Field Stations (OBFS; http://www.obfs.org/), an association of more than 200 *field stations* and professionals concerned with *field* facilities for *biological* research and education. The survey's intent was to ascertain stations' general cyberinfrastructure needs, and specifically how the Jetstream cloud could server their needs. The results of this survey are being analyzed, but an initial result is illustrated in Fig 4. After more analysis, we will generate and make availible a white paper reporting the results of this survey.

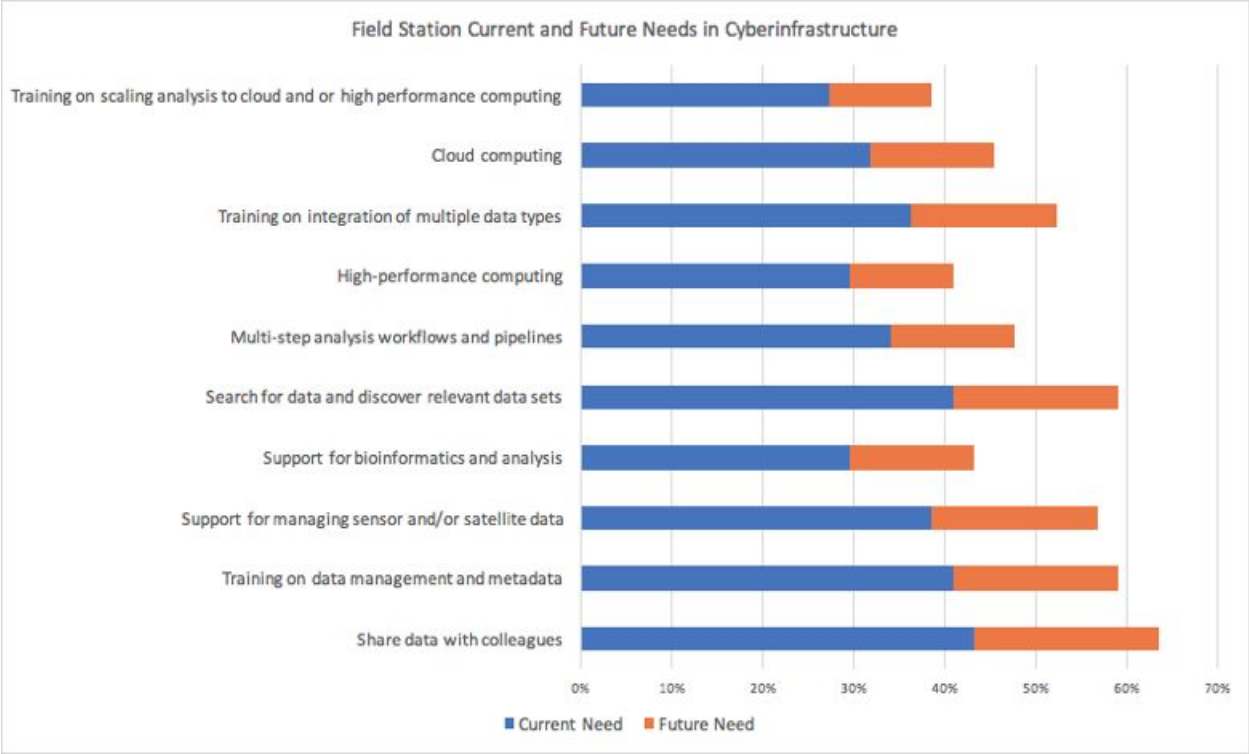


Figure 4. Cyberinfrastucture needs of field and marine stations

Synergistic activities

NCGAS is both a specific NSF-funded service provider in genomics, and a management group in IU's Pervasive Technology Institute. In this second guise, the NCGAS takes part in other projects that we feel augment our NSF-funded services.

- Active engagement with the Galaxy development community. NCGAS and IU hosted last year's Galaxy conference (~300 participants) and NCGAS member Carrie Ganote was on the organizing committee and a presenter, and is now on the finance committee.
- Active engagement with the Generic Model Organism Database (GMOD) community. This is a work in progress, but as we invest effort in genome browsers we hope to play a role in GMOD activities. We are collaborating with professor Naomi Stover at Bradley University, who maintains several ciliate browsers.
- NIH ITCR-funded Trinity development and Galaxy hosting. Involvement of NCGAS and the IU Scientific Applications and Performance Tuning group has both improved Trinity and made it far more available. While aimed at cancer research, Trinity is extensively used by our non-medical clients, esp. where obtaining a genome assembly is not feasible (e.g. marine copepods and polyploid salamanders). Thus, Trinity *de novo* assemblies are most useful for the least "model" of our users' organisms. The IU Trinity Galaxy has 775 registered users, and gives NCGAS a global reach (Fig. 3).

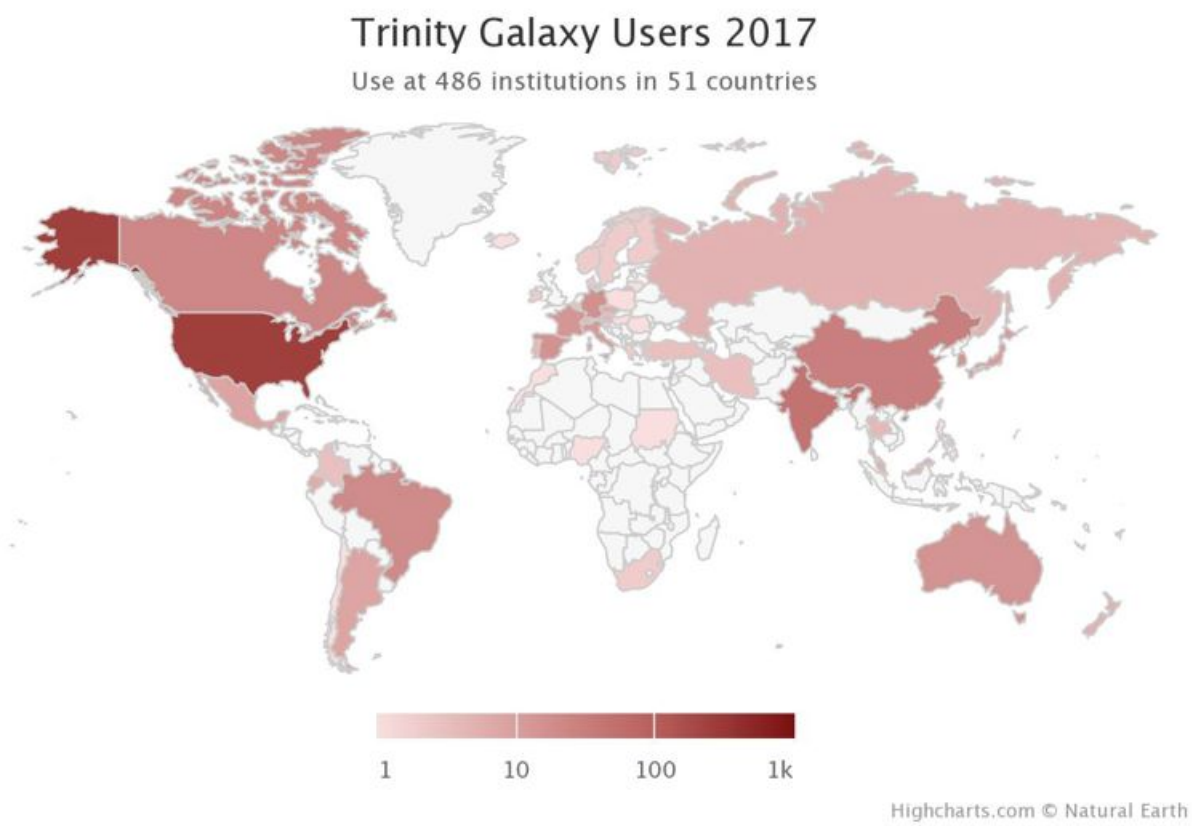


Figure 5. Users of the NCGAS-supported Trinity Galaxy instance worldwide

- NIH/NSF/ITCR-funded GenePattern hosting. Similar to Trinity, hosting GenePattern gives us an understanding of alternative software, and makes them available to our users.
- Keithanne Mockaitis specializes in plant transcriptomics, and as such has been and is a member of many national and international consortiums characterizing commercial plants, including mango, cocoa, and loblolly pine. Current projects are coffee, peanut, and sweet potato. Both Ganote and Sanders are paid from NSF and USDA grants Mockaitis is a co-PI on.

This mix of activities provides a diversified funding base to the NCGAS management group.

Figure 4. In addition to the ABI funding to support NSF researchers, the IU NCGAS management group undertakes a number of other synergistic projects, supporting by other funding sources.

What is the impact on the development of the principal discipline(s) of the project?

Results have been disseminated to communities of interest in a variety of ways, including:

- Publications in scientific journals
- Presentations

- Birds of a feather sessions at technical conferences
- Displays and booths at national and international technical conferences
- Articles in the lay press, most notably in Science Node (formerly International Science Grid this week - now at https://sciencenode.org
- NCGAS web site at ncgas.org
- In-person contacts
- Email list distribution
- Newsletter

What is the impact on other disciplines?

The primary other discipline on which NCGAS has had an impact is computational science and cyberinfrastructure. The largest impact that NCGAS has had in computational science has been to establish a model of a "domain-specific scientific service center", independent of federally-funded cyberinfrastructure computational resources; we have decoupled federal funding for supercomputers and funding for supercomputer application support. This ensures that a community relatively new to supercomputers use —biology for example—has support funded by the BIO directorate of NSF and is attuned to the needs of current research in the field.

NCGAS has just become an XSEDE Domain Champion, a new category supplementing Campus Champions, who are domain agnostic. As a very recent XSEDE program, we will see how Campus Champions works out.

We have also established new models for distribution of software relevant to biological research, which improves the nation's ability to use its aggregate cyberinfrastructure resources.

What is the impact on institutional resources that form infrastructure?

The software distributed by NCGAS has improved the effectiveness and ease of use of cyberinfrastructure resources throughout the nation.

What is the impact on information resources that form infrastructure?

NCGAS has facilitated the publication of several data sets important to basic biological research and to management of important plant and animal stocks. In the future, NCGAS will place a greater emphasis on genome browsers, an important product of 'omic research.

What is the impact on technology transfer?

The primary impact of NCGAS on technology transfer is in providing a collection of genomics applications easily available to any researcher. In the case of Trinity and GenePattern NCGAS stands at the interface of developers and users.

What is the impact on society beyond science and technology?

The societal impact of genomic characterization is gradual, but can be tremendous over time. Even the human genome's impact was mutated at first and is still being explored. We can expect that understanding the genome of pine tree, cacao, and mango will allow these important crop plants to be better managed over coming decades. The potential impact of science supported by NCGAS on society through better management of food supplies and better understanding of how organisms adapt to global climate change could be of fundamental importance to US and global populations. The speed with which human microbiome characterization has both begun to inform medical decisions (in nearly every field of medicine, including cancer), and swept through popular media, is amazing.

Research Technologies Division of UITS

The mission of the Research Technologies division of UITS is to develop, deliver, and support advanced technology solutions that enable new possibilities in research, scholarly endeavors, and creative activity at Indiana University and beyond – and to complement this with education and technology translation activities to improve the quality of life of people in Indiana, the nation, and the world. As such, it is more of an engineering, delivery, and service organization than it is an innovation organization. However, pursuing excellent service for the IU community often involves innovating. A few key innovative projects led by Research Technologies during FY 2017 are described below.

- **Jetstream**. Jetstream is the National Science Foundation's first production cloud (NSF grant # 1445604). Jetstream was accepted as a production system by the National Science Foundation in FY 2016 and will undergo a system review by the NSF in July 2017. Jetstream is a first-of-a-kind system; it is a configurable large-scale cloud computing resource that leverages both on-demand and persistent virtual machine technology to support a much wider array of software environments and services than current national cyberinfrastructure (National Science Foundation supported) resources can accommodate. As a fully configurable cloud resource, Jetstream bridges a major gap in the current XD ecosystem, which has machines targeted at large-scale, high performance computing, high-memory, large-data, high-throughput, and visualization resources. In particular, Jetstream:
    - Provides "self-serve" academic cloud services, enabling researchers or students to select a virtual machine (VM) image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing.
    - Hosts persistent VMs to provide services beyond the command-line interface for science gateways and other science services. Galaxy will be one of the initial science gateways supported.
    - Enables new modes of sharing computations, data, and reproducibility. Jetstream will support Digital Object Identifier (DOI)-based publication and sharing of VMs via Indiana University's persistent digital repository, IUScholarWorks, as well as supporting all Globus services, including data transfer/sharing with Globus Connect, and identity federation through Globus Nexus.
    - Expands access to the NSF XSEDE ecosystem by making virtual desktop services accessible from institutions with limited resources, such as small schools, schools in Experimental Program to Stimulate Competitive Research (EPSCoR) states, and Minority Serving Institutions. For example, VMs will enable use of Linux desktops from tablets.

- **Programmable Immersive Peripheral Environment Systems (PIPES)**. PIPES is a tool that extends commonly used virtual reality systems in support cyber-physical applications. PIPES allows a developer to programmatically control high voltage sockets. Students and staff have used PIPES in virtual environments to simulate wind, heat, and scent conditions thereby improving the

immersive quality and increasing their suspension of disbelief. PIPES was developed by Chauncey Frend from the Advanced Visualization Lab (AVL) and was awarded the " 2016 Best Research Demo Award" by the IEEE Computer Society (2016 IEEE Virtual Reality conference in Greenville, SC) and portions of PIPES technology are patent-pending.

- **Open XD Metrics on Demand Value Analytics (XDMoD VA)**. XDMoD VA is a project funded by the National Science Foundation (NSF grant #1566393). Understanding the value of campus-based cyberinfrastructure (CI) to the institutions that invest in such CI is intrinsically difficult. Given today's financial pressure, administrative support for campus-based CI centers offering resources to local campus users is under constant budgetary pressure. This is partly due to the difficulty in obtaining quantitative metrics that clearly demonstrate the utility of investment in campus CI centers in enhancing scientific research and the financial aspects of enhanced competitive ability in seeking funding for research. This proposal seeks to implement a set of highly experimental modules to be added to the existing CI metrics tool eXtreme Data Metrics on Demand (XDMoD). These modules are intended to provide a means for assessing campus-based CI investment in scientific terms, as measured in publications, and in financial terms, as measured in grant income from researchers who use CI as compared to those who do not. The modules, tentatively called Open XDMoD Value Analytics, will present a view of financial, collaboration, and publication data, showing "return-on-investment" metrics in relation to CI usage. The goal of the work is to generalize these modules to work with a variety of funding and publication information sources, and incorporate them into the open source distribution of Open XDMoD.
- **Carbonate**. The new IU compute resource Carbonate is available to users through one.iu.edu starting July 2017. Carbonate consists of 96 nodes each w/24 CPU cores and 256 GB of RAM. It will serve as a replacement to the Mason cluster (8 of the 96 nodes have 512 GB of RAM), provide additional research desktop services (12 additional desktop nodes), and general purpose larger memory nodes (2 for login services, and 74 batch computing).
- **Karst Desktop**. Karst Desktop is a remote desktop service for users with accounts on the Karst research supercomputer. Karst Desktop lets researchers open and control a remote session on Karst from a graphical desktop window running on your personal computer. A researcher can run graphical applications installed on Karst from your Karst Desktop window without noticeable latency. Additionally, a researcher can export your computer's local drives and directories, making them available to you during your remote session, simplifying the transfer of data between your computer and your Karst account. Overall, Karst Desktop provides numerous features that are especially helpful to users who are unaccustomed to working in Unix-like command-line environments. Today, we already have over 700 unique users that have used the service and over 250 unique users login every month. https://kb.iu.edu/d/bfwp
- **Open Science Grid.** IU participates as a funded member of the NSF-funded Open Science Grid project (NSF grant #1148698). The basic idea of Grid Computing is to utilize available CPU cycles and storage of many computer systems across a worldwide network so that they can function as a flexible, pervasive, and inexpensive accessible pool that could be harnessed by an individual, accredited user, similar to the way power companies and their users share the electrical grid. The Open Science Grid (OSG) is the major facilitator of Grid Computing in the U.S. Researchers subsequently developed these ideas in many other exciting ways, producing for

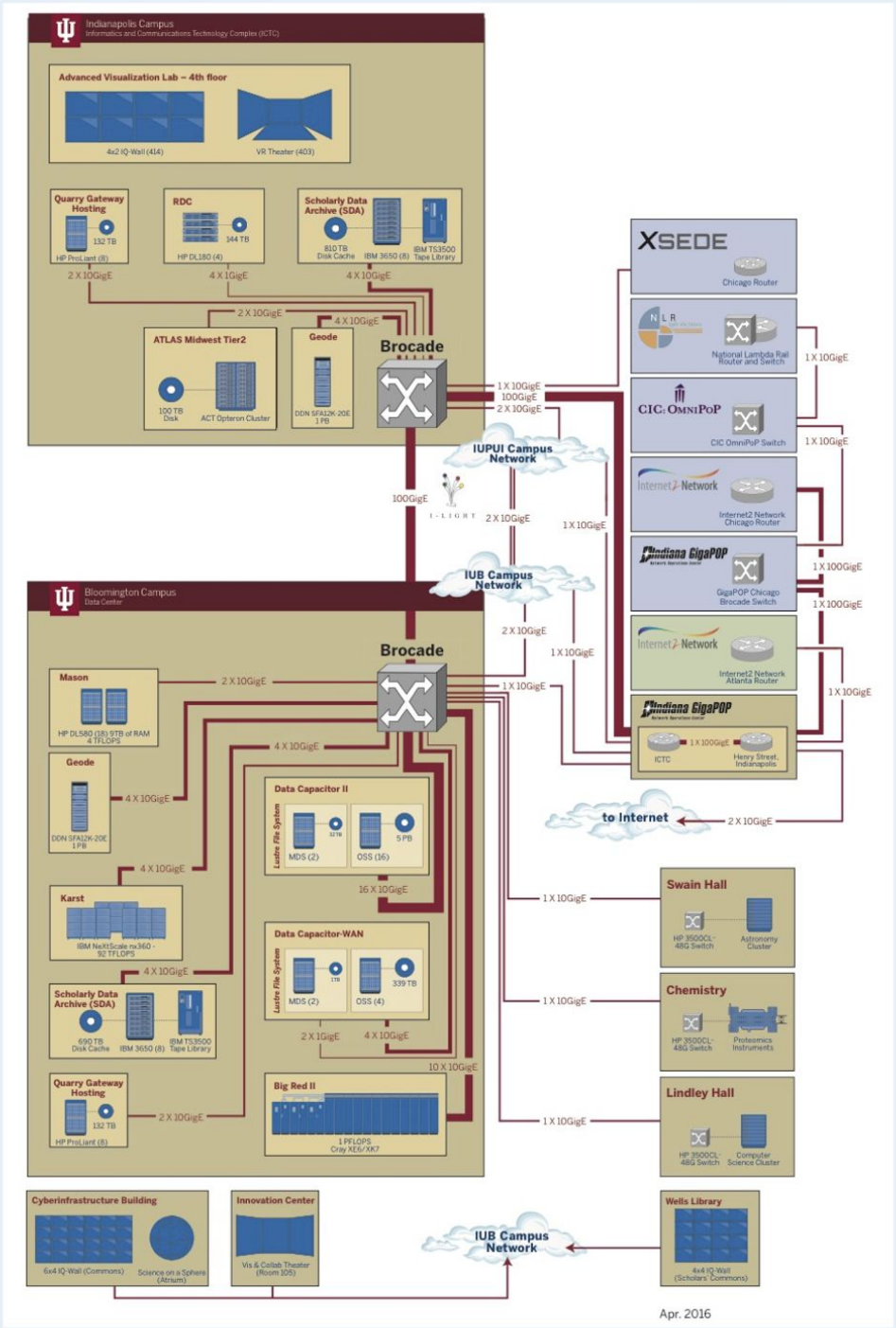
example, in addition to OSG, large-scale federated systems (TeraGrid, EGEE, Earth System Grid) that provide not just computing power, but also data and software on demand.

- **CODATA-RDA School of Research Data Science.** RT is a proud participant in the Research Data Alliance (RDA). The CODATA-RDA School of Research Data Science has developed a short course, summer school, style curriculum that addresses training requirements for research data management and data management plans. The School maintains an international network of these schools, some of which are taught by staff from RT. For more information on the schools, see: http://www.codata.org/working-groups/research-data-science-summer-schools
- **Photogrammetry**. RT has 60 licenses for PhotoScan that can be used in parallel to create 3D models and run from Karst Desktop so that the advantages of PhotoScan's GUI are not lost. In the past, stitching together thousands of photos for these models could take weeks. However, by leveraging the power of high performance computing systems like Karst and Carbonate runtimes decrease dramatically. Users come from disciplines such as anthropology, archaeology, art history, and informatics.
- **Text Analysis.** RT has developed workflows in R and Python for basic text analysis which are available on github https://github.com/cyberdh/Text-Analysis and used as introductory material to the languages for novice users. More sophisticated algorithms and workflows such as topic modeling and tailored sentiment analysis are available in consultation with CyberDH.
- **Reality Labs**. In collaboration with Learning Spaces, the AVL has deployed Reality Labs in classrooms at IU. Reality Labs are created from existing classroom or lab spaces and contain an instructor station plus around 10 to 30 student stations. Each of these stations includes an Acer G1 Predator gaming computer and an HTC Vive head-mounted VR display. A shared, room-wide VR tracking space facilitates immersion in the virtual simulations. At the same time, a large-format display (such as a projector or tiled video wall) allows non-immersed users to share in the virtual experience and also supports collaborative critiques and reviews. Several articles have been written about Reality Labs:
    - https://campustechnology.com/articles/2017/03/07/when-virtual-reality-meets-the-classroom.aspx
    - http://www.magbloom.com/2017/06/ius-reality-lab-offers-students-an-out-of-this-world-experience/
    - http://mediaschool.indiana.edu/news/schools-virtual-reality-lab-available-for-variety-of-uses/
    - http://archive.inside.indiana.edu/features/stories/2017-02-22-virtual-reality.shtml

Additional highlights of RT systems, services, capabilities, and activities include:

Highlight: Overview of RT Cyberinfrastructure

Schematic diagram of IU cyberinfrastructure showing network connections between IU and other national networks and network connections and cyberinfrastructure within IU.



Highlight: How damaging were those 2015 Indonesian forest fires?

When widespread forest and peatland fires broke out across Indonesia in fall 2015, scientists knew the air pollution would have a significant effect on human health.

To figure out the extent of the damage, though, researchers relied on the innovative high performance computing and storage resources at Indiana University. In a paper published in **Scientific Reports** in late 2016, lead author Paola Crippa, from Newcastle University, United Kingdom, shows that, in fact, the fires exposed 69 million people to unhealthy air pollution.

How did she come up with that staggering number? She and her fellow researchers from the United Kingdom, United States, Singapore, and Indonesia used a high-resolution air-quality model to simulate the impact of the fires on air pollution. All of the simulations and models were analyzed and are now stored on high performance computing and storage resources at IU, including the supercomputer Big Red II, the Scholarly Data Archive, and Data Capacitor II. But what, exactly, brought together a European researcher and IU's HPC tools? It turns out that Crippa, who earned her doctorate at IU, is well acquainted with IU's supercomputing prowess. She already had a strong working relationship with Abhinav Thota, a principal research software engineer in IU Research Technologies, and she knew that he and IU could handle the massive data analysis and storage she needed.

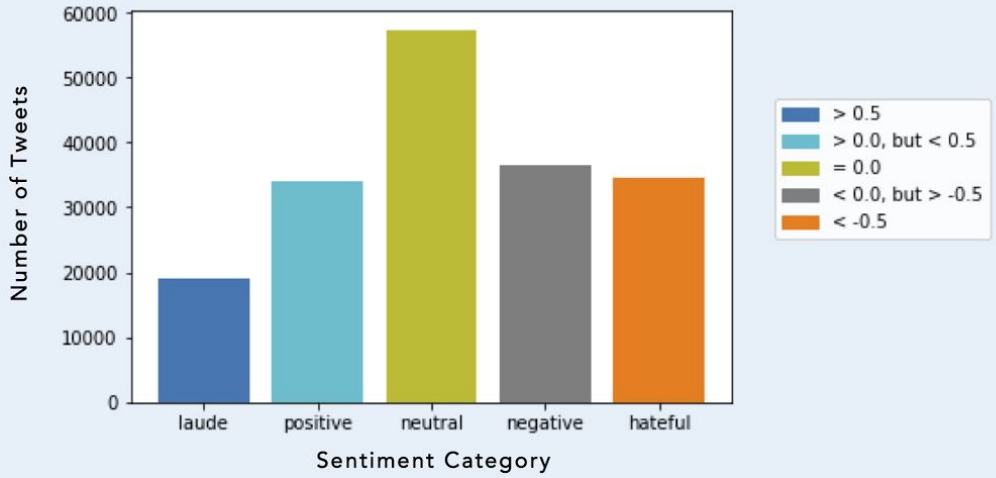
In true Hoosier fashion, Thota was happy to help. "This was a really rewarding project for me," he said. "Providing Big Red II to the researchers allowed them to do productive, important science with broad implications for society at large." For his part, Thota is named a co-author on the paper. The original story appeared here.

Abhinav Thota working with Big Red II.

Highlight: Sentiment about the POTUS (100 days)

Since FDR mentioned the first 100 days in a speech in 1933 and then passed 15 major bills in the next 105 days to help dig America out of the Great Depression, 100 days has become a yardstick to measure incoming presidents. It is generally considered the time when an incoming president has the greatest power and influence and how they use said power and influence determines their ability to govern. The Cyberinfrastructure for Digital Humanities group, Research Technologies at IU, assessed the public's sentiment about President Trump by performing *sentiment analysis* (a method for analyzing the feeling – positive/negative/neutral – of text in a systematic fashion) on tweets from the first 100 days of his presidency.

Using 181,803 tweets (from January 20th, 2017 to April 29th, 2017), CyberDH used the VADER[1] sentiment analysis. For the VADAR analysis, words, emojis, all caps letters, and even slang like "lol" or "smh" (shake my head) are scored based on an assigned positive or negative value – the final result gives numerical negative, neutral, and positive scores as wells as a final compound score between -1 and +1. In addition, "good" is not as positive as "great" so "great" is worth more points. The code can be found here, yielding the following results:



Final compound scores (ranging between -1 and +1) for the tweets about President Trump for the 1st 100 days.

This sentiment analysis shows that most of the tweets were *neutral*. This is an interesting finding because with the VADER analysis, it is difficult for one word to completely cancel out another meaning that tweets tend to be scored as either positive or negative. This finding is most likely the result of most of those *neutral* tweets not having any words or emojis or anything that indicated sentiment one way or another.

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Highlight: New workshop at SC16 conference

Jenett Tillotson, Research Technologies at Indiana University, along with colleagues from different universities organized the first HPC System Professionals Worshop at the SC16 (SuperComputing) conference. Highlights from the workshop included an invited paper as well as four competitive papers (all of which were published on IUScholarWorks).

Invited Paper: HPC Systems Acceptance: Controlled Chaos

Paper: Account Management on Large-Scale HPC by Brett Bode

Paper: Cluster Computing with OpenHPC by Karl Schulz

Paper: Increasing HPC Resiliency Leads to Greater Productivity by Roger Moye

Paper: Blue Waters Resource Management and Job Scheduling Best Practices by Jeremy Enos

Information | Invited Paper | Schedule | Topics | Information | Organizing Committee | Program Committee | Call for Papers | Contact | Links

Inaugural HPC Systems Professionals Workshop

HPCSYSPROS16

November 14, 2016, Salt Lake City, Utah

SC16 Salt Lake City, Utah hpc matters.

Held in conjunction with SC16.

QUICK INFORMATION

In order to meet the demands of high performance computing (HPC) researchers, large-scale computational and storage machines require many staff members who design, install, and maintain these systems. These HPC systems professionals include system engineers, system administrators, network administrators, storage administrators and operations staff all who face problems that are specific to high performance systems.

The Systems Professionals Workshop intends to be a platform for discussing the unique challenges that come from supporting large-scale, high performance systems. We are soliciting submissions that speak directly to the state of the practice of standing up and operating high performance systems with an emphasis on solutions that can be implemented by systems staff at other institutions.

We will be presenting the *Inaugural HPC Systems Professionals Workshop* at SC 16 on Monday afternoon (November 14, 2016).

INVITED PAPER: HPC SYSTEMS ACCEPTANCE: CONTROLLED CHAOS

System acceptance is critical when deploying HPC resources from small clusters to large supercomputers. Each type of system has its own challenges and having a well established and proven test, acceptance, and integration plan is an invaluable tool to the site and the vendor. The topic of systems acceptance is quite broad. This talk will be mostly focused on the system?s software and hardware components.

Biography:
Paul Peltz is a Scalable Systems Engineer in the High Performance Computing Division at Los Alamos National Laboratories where he helps procure, test, and integrate the new HPC systems into production. He has almost 20 years of experience collaborating with vendors to evaluate and integrate pre-release hardware and software. Currently he is the systems technical lead on the Trinity project and is actively working with Cray to deploy their largest XC system to date, a 110 cabinet XC40 named Trinity.

Screenshot from HPCSYSPROS (SC16) conference website

According to Tillotson, HPCSYSPROS16 went so well that the group is planning on expanding efforts for SC17 which will be held in Denver, CO in November 2017.

Appendix 1: EOT Activities

EOT activities during FY 2017				
Event	Education, Outreach, and Training Event Title	Conference Name/Location	Description	Total Attendees
7/5/2016	AVL Tour	STEM girls camp / IUPUI - Informatics & Communications Technology Complex AVL	Informal tour of AVL systems and IU research projects.	27
7/8/2016	Funding models initiatives and approaches	CASC/UK HPC-SIG Workshop / Oxford University e-Research Centre	Discussion with members of the US and UK HPC community	16
7/11/2016	How to create virtual reality applications	SOIC Workshop - Gadgets of Virtual Reality / IUPUI - ICTC IT 419	5 day workshop where AVL staff did the teaching.	8
7/13/2016	AVL Tour for NEST teachers	IUPUI - Informatics & Communications Technology Complex AVL	Informal AVL tour	11
7/14/2016	What's new with the Indiana Spatial Data Portal	IUB - Herman B Wells Library	Overview of new functionality of RT ISDP service	8
7/15/2016	Advanced Visualizaiton support	Indiana Black Expo / Indianapolis, IN	AVL helped at IN Black Expo	71
7/18/2016	Hands on with Jetstream	XSEDE16 Miami FL	Tutorial for using Jetstream via Atmosphere	16
7/20/2016	SOIC Workshop AVL Tour July 20th	IUPUI - Informatics & Communications Technology Complex AVL	Informal tour of AVL systems lead by AVL Staff	6
7/21/2016	Hands on with Jetstream - ECSS API Tutorial	XSEDE16 Miami FL	Jetstream API training for ECSS/Staff	35
7/27/2016	BioInformatics Summer Workshop AVL Tour	SOIC BioInformatics Summer Workshop / IUPUI - Informatics & Communications Technology Complex AVL	Informal tour led by AVL staff and SOIC instructors.	20

7/27/2016	2D Animation Summer Workshop AVL Tour	SOIC 2D Animation Summer Workshop / IUPUI - Informatics & Communications Technology Complex AVL	Informal tour led by AVL staff.	6
7/31/2016	Environmental Genomics Workshop	MDIBL / Salisbury Cove Maine	NCGAS provided HPC support for genomic analysis training on how to use cluster resources and command line as well as genomic analysis support	18
8/2/2016	AVL Tour for SCORE rep. Mike Crumbo	IUPUI - Informatics & Communications Technology Complex AVL	Informal tour led by AVL staff.	1
8/3/2016	AVL Tour for Pakistan Education USA (SOIC Guests)	IUPUI - Informatics & Communications Technology Complex AVL	Informal AVL tour	29
8/17/2016	AVL Tour for IUPUI Admissions as Guests of SOIC	IUPUI - Informatics & Communications Technology Complex AVL	Informal tour to show and demonstrate what SOIC students do in the AVL	18
8/17/2016	AVL Tour for SOIC Business Guests	IUPUI - Informatics & Communications Technology Complex AVL	Informal tour of the AVL from the perspective of why SOIC researchers use the lab.	3
8/17/2016	Jetstream Overview	Purdue University	Presented on Jetstream to researchers and IT staff	14
8/18/2016	AVL tour for PCR Group (SOIC Guests)	IUPUI - Informatics & Communications Technology Complex AVL	Informal tour led by AVL staff.	45
9/2/2016	NCGAS: HPC/Bioinformatic support they provide to various groups	Teleconference organized by NIH/NCI	NCGAS presented to PIs of the NIH ITCR study group	48
9/9/2016	AVL overview presentation	CyberDH/AVL workshop series / Wells Library IUB	presentation	16
9/13/2016	AVL Tour for JagDays Guests	IUPUI - ICTC Advanced Visualization Lab	Informal lab tour lead by AVL staff.	15
9/13/2016	Campus Visit for leaders from University of South Dakota	IUB - Cyberinfrastructure Building	Informal tours and discussion	4

9/22/2016	AVL Tour for Pike HS iDEW (SOIC guests)	IUPUI - ICTC Advanced Visualization Lab	Series of tour rotations lead by AVL staff and SOIC students and staff.	109
9/23/2016	AVL Tour for Arsenal Tech HS iDEW (SOIC guests)	IUPUI - ICTC Advanced Visualization Lab	Hosted by AVL staff and SOIC staff and students.	41
9/23/2016	Sustainability of Coffee Production in Colombia: The National Colombian Coffee Growers Federation its National Coffee Research	IUB - Cyberinfrastructure Building IQ-Wall	NCGAS organized a meeting between major coffee research group Cenicafe and collaborators at Cornell	10
TOTAL				**595**